PB-239 721

STATISTICAL CONCEPTS FOR DESIGN ENGINEERS

J. R. Murphy, et al

Oklahoma State University

Prepared for:

National Environmental Research Center

September 1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-650/2-74-080	2.	3. PB 239 721
4. TITLE AND SUBTITLE Statistical Concepts for Design Engineers		5. REPORT DATE September 1974 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) J. R. Murphy and L. D. Broemeling		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Oklahoma State University Stillwater, Oklahoma 74074		10. PROGRAM ELEMENT NO. 1AB013; ROAP 21ADE-026 11. CONTRACT/GRANT NO. Grant R-802269
12. SPONSORING AGENCY NAME AND ADDRESS EPA, Office of Research and Development NERC-RTP, Control Systems Laboratory Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED Final; Through 8/74 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The report describes basic statistical concepts for engineers engaged in test design. Although written in handbook form for use within the Environmental Protection Agency, it is not intended to replace existing statistics textbooks. Its objectives are: to enable design engineers to converse with consulting statisticians, to introduce basic ideas for further individual study, and to enable the reader to make some immediate applications to his own work.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

PRICES SUBJECT TO CHANGE

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Statistics		12A
Statistical Inference		
Experimental Design		14B
Environmental Engineering		05E
Air Pollution		13B

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 152
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE $6.25-2.25

EPA Form 2220-1 (9-73)

EPA-650/2-74-080

STATISTICAL CONCEPTS FOR DESIGN ENGINEERS

by

J. R. Murphy and L. D. Broemeling

Oklahoma State University
Stillwater, Oklahoma 74074

Grant No. R-802269
ROAP No. 21ADE-026
Program Element No. 1AB013

EPA Project Officer: W. R. Schofield

Control Systems Laboratory
National Environmental Research Center
Research Triangle Park, North Carolina 27711

Prepared for

OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

September 1974

This report has been reviewed by the Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

TABLE OF CONTENTS

Chapter Page

0. PREFACE vi

1. INTRODUCTION 1

2. POPULATIONS, VARIABILITY, UNCERTAINTY, AND SAMPLING 5

3. VARIABILITY AND RANDOM ERROR 14

4. BASIC CONCEPTS OF PROBABILITY AND MATHEMATICAL STATISTICS . . . 19

  4.1 Random Experiments and Probability 19

  4.2 Random Variables and Distributions 24

  4.3 Moments and Expectation 32

  4.4 Other Descriptive Quantities 35

  4.5 Jointly Distributed Random Variables 37

  4.6 Useful Distributions 41

    4.6.1 The Binomial Distribution 42

    4.6.2 The Poisson Distribution 44

    4.6.3 The Geometric Distribution 45

    4.6.4 The Exponential Distribution 46

    4.6.5 The Normal Distribution 47

    4.6.6 The Chi-Square Distribution 52

    4.6.7 Student's t-Distribution 52

    4.6.8 The F-Distribution 53

5. SAMPLING AND INFERENCE 54

  5.1 Description of Finite Number Populations 54

  5.2 Statistical Inference in Finite Populations 55

  5.3 Statistical Inference in Infinite Populations 66

  5.4 Statistical Inference in Normally Distributed Populations . . 70

Page

5.4.1 A Single Normal Population 71
5.4.2 Two Non-Independent Normal Populations 78
5.4.3 Two Independent Normal Populations 81
5.4.4 Several Independent Normal Populations 85
5.4.5 Analysis of Variance 93
6. EXPERIMENTAL TEST PROGRAMS 99
6.1 Introduction 99
6.2 Data 100
6.3 Comparative Experiments 104
6.4 The Use of the Word "Design" 105
6.5 Properties of a Good Experiment 106
6.6 Experimental Units 109
6.7 Experimental Error and Sampling Error 110
6.8 Degrees of Freedom in the Analysis of Variance 114
6.9 Randomization 115
6.10 Randomized Designs 118
6.11 Multifactor Experiments 120
6.12 Mathematical Models 127
7. SUMMARY 136
8. REFERENCES 138

FIGURES

No. Page

1. Binomial Probability Mass Function 43
2. The Normal Density Curve 47
3. Binomial Mass Function 50
4. Results of the 0.05-Level LSD Test ($LSD_{0.05} = 12.1$) 91
5. Experimental Design 105

0. PREFACE

In the last few years, data gathering and analysis has reached such a level of sophistication that in many cases, statistical treatment is considered essential. As a result, many scientists and engineers have become acquainted with the valuable assistance that statistical methods can often provide. In addition, they have helped to point out where new and/or better techniques are needed, so that statistical theory and statistical techniques have also reached a high level of sophistication. Consequently, the study and practice of statistics, like many disciplines, has become a specialty field, which has tended to cause some science and engineering to feel uncomfortable about the subject and to regard it too difficult to understand and/or utilize. To be sure, statistics is not a trivial subject, and the effective application of statistical methods to real world problems does require special training that an engineer oft-times simply cannot afford to undertake. Thus, the business of designing and conducting experiments, and analyzing and interpreting the results has necessarily become a joint effort requiring teamwork between specialists. It is a common misconception that such teamwork is of a "production assembly line" nature, where each member does his part and then passes it on to the next. Thus, many times, a consulting statistician is not introduced to a project until the data is already gathered and it is time to analyze and interpret it. It would seem that the researcher who has invested his time in planning and carefully conducting an experiment would be reluctant to entrust its analysis and interpretation to one who, although an expert in techniques of analysis, is relatively unfamiliar with the circumstances surrounding the data; however, there are many who believe that this is what they are supposed to do. Perhaps much of the distrust some people have for "statistics" is attributable to such a misconception.

One of the points we shall try to emphasize in this manual is that useful analysis of data cannot be based on the numbers alone, that in order to be able to draw meaningful inferences from the results of an experiment, one must consider how those results were obtained. Thus, when we speak of teamwork, we are referring to cooperation and interaction of the members at all stages of a project. If a statistician is to be a member of the team, there are at least two reasons why he should be included from the start. (1) In line with the thought expressed above, a great deal of time can be saved. Since the analysis of the data depends so much upon the background, invariably a statistician brought in at the later stages must be filled in and brought up to date. On the other hand, if he is already familiar with the project, the lag time between collection of the data and analysis of the data is less and, in addition, the analysis is more efficient (2) More importantly (as we shall also try to show throughout the manual), only a small part of statistical training is concerned with data manipulation. "Statistics" is to a large degree a way of thinking based upon principles general enough to allow application to a wide variety of special circumstances. As such, statistical thinking can make contributions to every phase of experimentation; sometimes the greatest contributions are made in the initial stages, say, in helping to recognize and define the problem.

We have emphasized that imaginative and timely analysis of data requires that the consulting statistician have at least a surface understanding of the field of knowledge pertaining to the experiment which gives rise to the data, and within which the results are to be interpreted.

A similar requirement also holds for the scientist or engineer who wishes to make effective use of the services of a statistical consultant. The best possible solution would be for one to be expert in both his own field and statistics as well, but such a capability is possessed by few individuals. An alternative is for the statistician to understand something of the field where his specialty is to be applied and for the engineer to be acquainted with the principles upon which the statistical treatment is based.

There are two erroneous ideas that people sometimes have about statistics that we hope to address throughout this manual. The first is that statistics is nothing more than a routine set of calculations to be performed upon numbers. Such a misconception is probably responsible for much of the inappropriate application of statistics that one can see today. Many people, noting that many statistical techniques involve the same standard calculations, get the idea that this is all there is to statistics, and that the same thing is applicable regardless of the situation. Then, there are those who acquire the opposite idea about statistics. Frequently, people without a strong mathematical background let themselves become "snowed" by the details of probability theory and mathematical statistics, and sometimes even by the underlying mathematics of applied statistics. A common attitude in such cases is that statistics is just so much mathematical "wizardry." It is not difficult to understand the existence of such misconceptions about statistics, because the fact is that statistical techniques do involve some routine calculations and are based upon an underlying mathematical structure that is sometimes detailed and difficult for someone not grounded in mathematics to easily grasp.

It will be our objective to present the reader with an overview of statistics as it relates to the planning, analysis, and interpretation of scientific experimentation. We shall not dwell on convincing anyone of a need to use statistics; the increasing evidence of use (and misuse) of statistics in scientific and engineering literature attests to the fact that engineers who shun statistical methods entirely are fast becoming a minority. Statistics has a valid case to make, but it has been well presented elsewhere. The manual is not intended to be a text nor a complete exposition of statistical methodology, and no detailed treatment of any of the subjects covered will be given. There are many very good references available on engineering and industrial statistics in which one may find detailed information, and the creation of yet another is beyond the scope of our limited objectives. It is our belief that it is possible for anyone trained in technical and scientific subject areas to understand the basic principles upon which statistical methods are based without having to wade through all the details, and the presentation will be largely limited to fundamental concepts. We shall assume that the reader is relatively unfamiliar with statistics and has neither the time nor the inclination to become expert in the field. It is anticipated that the manual will provide sufficient background so that the project engineers to whom it is addressed will have the ability to:

i. Know what statistics can and cannot do.

ii. Recognize potential areas for the application of statistical methods.

iii. Aid in the direction and thrust of statistical analysis of data.

iv. Share a greater role in the interpretation of results and the formulation of conclusions and recommendations.

v. Make efficient use of whatever statistical consulting expertise is available to them.

For the reader who may be interested in delving further into matters we may touch upon, a list of references is provided, many of which are considered classics in their respective subject areas.

1. INTRODUCTION

The area of statistical theory that this manual will primarily deal with is commonly referred to as inference. If statistics, as viewed from this angle, can be summarized in one statement, that statement must surely be "The role and commission of statistics is to aid the researcher in drawing optimum inferences from his data in the face of uncertainty." Sounds very good, but....

First, "to aid" means precisely that. There is no method or technique to replace the use of common sense. Indeed, there are those who would claim that application of statistical principles is merely a structured application of common sense. Statistics is not a mystical collection of hocus-pocus to be applied to the data, suddenly making every hidden secret become clear. Statistical inference, as an entity separate and distinct from scientific inference, does not exist. Statistical methods and techniques are not things "better left to the statisticians;" rather they are tools to be used in conjunction with and consistent with the basic principles of scientific inference. Robert Hooke in his monograph, Introduction to Scientific Inference, p. 94, offers this observation, "There is an old story about a flea trainer who claimed that fleas hear with their legs. As proof, he taught some fleas to jump at his shout of 'Jump!' After amputating the fleas' legs and observing that they no longer responded to his shouts, he rested his case. Statistics, of course, is an aid to, not a substitute for, intelligence. The flea trainer could have made measurements of great precision and gathered pages of data, but these would not have protected him from his faulty logic."

The word "optimum" requires explanation and qualification. To one who has not thought carefully about the meaning, "optimum" often carries the idea of the best, in all possible senses of being best. In actuality, when the word is used in an precise way, it means best only according to a specific criterion of goodness. Thus, an optimal procedure is optimal (in a real sense) only if the criterion is relevant.

Many optimality criteria used in statistics are based upon conceptual long-run properties. It should be clearly understood that these "in the long-run" or "on the average" properties cannot be imparted to a single trial or single application of a statistical method. It may be that such long-run properties will not seem terribly compelling to a researcher who has before him a single experiment, and he may wish for something more. But neither wishing nor cleverly devised words can change a long-run property into a short-run property.

Finally, the phrase "in the face of uncertainty" must be clarified. One of the basic premises of statistics is that observable phenomena are subject to variation and, although some of the causes or sources of variation can be accounted for, ultimately there remains variation due to unknown and unexplained sources. Frequently the assumption is made that the behavior of such variation can be described in terms of chance or random occurrence. That is, an exact deductive mathematical theory, the theory of probability, is used to define a systematic structure, within which variation can be placed and studied. The use of mathematical models for studying and understanding the physical world is not a new development by any means, but the idea that unsystematic chaos, disorder, or chance happenings can also be understood in the context of a logical system is a relatively new one. The consequences of using probability models for studying variability are far-reaching and, sometimes, almost astounding. Many statistical techniques that have arisen as a result appear to be very powerful, almost like "getting something for nothing." However, we must remind ourselves that there is no way to "create knowledge" with any statistical technique. Assumptions are made in the absence of deffinite knowledge and, no matter how reasonable, plausible, or compelling the assumptions may be, the fact remains that derived methods are based on and tied to those assumptions.

It is reasonable to ask how statistical theory and methodology may be utilized by engineers "to aid in drawing optimum inferences" from their data. Part of the answer comes with the realization that statistics has something to contribute to every phase of experimentation; that, in fact, the greater contribution may be in obtaining the data rather than analyzing it after it has been gathered. It may be useful to consider alternate descriptions of the use of statistical methods in experimentation which have been given by various engineering people who apply them:

1. With respect to experimental planning and data taking,
    (a) Statistical methods are aids in planning orderly experimentation.
    (b) Statistical methods help one to get the most information for the least amount of experimentation.
    (c) Statistical methods help one to organize, categorize, and quantify his data.

2. With respect to analysis of data,
    (a) Statistical methods provide ways of condensing and summarizing data with the minimum loss of information.
    (b) Statistical methods assist in determining what is apparently systematic and what is apparently random in a set of data.
    (c) Statistical methods enable one to get a "complete picture" of the way the relevant factors in an experiment are affecting the response of interest.
3. With respect to inferences and conclusions,
    (a) Statistical methods permit one to determine how far his results can be safely generalized.
    (b) Statistical methods give the user "yardsticks" by which the strength of his results may be measured.
    (c) Statistical methods provide a quantification of the degree of uncertainty associated with estimates made from the data.

The list is, of course, incomplete. This manual, it is hoped, will be useful in assisting engineers to avail themselves of the contributions that statistical methods can make.

2. POPULATIONS, VARIABILITY, UNCERTAINTY, AND SAMPLING

The field of statistics began with and has traditionally been considered as concerned with the gathering, organization, and analysis of facts in order to determine the essential characteristics of interest of some population under study without having to examine the population in its entirety. At first, the populations studied were largely existing ones such as people or items of agricultural or industrial production, and the main concerns were efficient sampling schemes and techniques of analysis which reduced the risk of erroneous inferences about the population when generalizing from the sample. Let us note here an evident fact, and that is that any study of a population of physical objects necessarily involves some form of measurement of one or more attributes of interest. Thus, any given population of objects can give rise to several number populations, depending upon what aspect of the physical population we wish to study, and it is common usage to speak of "the population" in reference to either the physical population itself or to some number population derived from it by measurement. (We include here measurements of the classification type, or so-called qualitative measurements such as "present" or "not present," or such as "belongs to Class 1, Class 2,..., Class n," because these measurements can be represented as numerical measurements. For example, we can use the correspondence, 0 = "not present" and 1 = "present." Hence, we shall think of all measurements of populations as being numerical measurements.)

As the ideas of statistics continued to develop, it became apparent that the principles used in the study of existent populations were also applicable for studying conceptual populations, that is, populations which did not exist already, but could be theoretically generated if necessary. For example, we may conceive of the population generated by one million tosses of a coin, the population generated by a large indefinite number of rolls of a pair of dice, or the population of responses generated by repeated measurement of some object or substance. Conceptual populations may be finite or infinite. To carry the notion further, from a given population, conceptual or otherwise, it is conceptually possible to generate a derived population of samples (actually, several populations, one for each fixed sample size) by repeatedly drawing samples from the original (parent) population. Conceptual populations generated by repeated sampling are very important types of populations, and are the kind a large portion of future discussion will be concerned with. From a statistical standpoint, in the study of any population, there are actually two types of populations involved, the parent population or the population of inference and one or more derived populations generated by repeated sampling. An important special case of this idea is a population generated by conceptual repetitions of an experiment. The study of the sampling techniques for populations to be studied by experimentation comprises a portion of the subject matter of the area of statistics known as experimental design. That is, the performing of an experiment can be thought of as taking a sample from some population, and the theory of experimental design is concerned, not only with the method of taking the sample and the population of inference (the experiment itself), but also with the population generated by repeated sampling (conceptual repetitions of the experiment).

Now, let us consider some of the concepts involved in the study of populations. First, there is the idea of variability, that is, that all members of the population do not give the same measured numerical response with respect to the attribute of interest. If there is no variation, there is no need of any statistical techniques, for the population will be known completely as soon as we pick some member and observe it. We shall take as a starting point, therefore, that the reason why any population requires more than examining just one of its members is that the population is subject to variation.

Assuming a population of interest does possess variability, it naturally follows that any conclusions about the population as a whole, based on observing a subset of the population, must be subject to uncertainty. Can the uncertainty be reduced? Yes, in fact it can be eliminated entirely if we can examine the entire population, but this is, of course, impractical in all but a few cases. The question arises as to whether it is possible to find a satisfactory measure of uncertainty in order to quantify the concept and give it more definite meaning. Consider, for a moment, some intuitive aspects of uncertainty: (1) the larger the subset of the population it is possible to examine, the less uncertain will be our conclusions; (2) the more variable the population is, the more uncertain our inferences must be; and (3) the more that we already know about the population, the less uncertain we are. Any measure of uncertainty should somehow embody these considerations.

In many cases, the reason for sampling can be taken to be the estimation of some quantity in the population, and this usually involves calculating some estimate from a sample. Focusing our attention on this sample estimate, we see that the derived population of samples gives rise to yet another population--the population of different values that the sample estimate may assume from one sample to the next. Due to variability in the parent population, the population of values of the sample estimate will also exhibit variability. Moreover, the variation of a sample estimate is generally governed by: (1) as the sample size increases, variation decreases; (2) the more variable the parent population, the greater the variation; and (3) with knowledge of the parent population, the sampling procedure can be modified in order to reduce variability. Thus, in terms of estimation, uncertainty may be measured in terms of the variability of the sample estimate over repeated sampling. A somewhat startling result, which we will later discuss in more detail, is that when sampling from a parent population is done in a certain way, the variability of a sample estimate can itself be estimated from the same sample. That means, that from a single sample, it is possible to get both an estimate of some population characteristic and a measure of the uncertainty of that estimate. Thus, although uncertainty can sometimes be reduced and sometimes not, it can and should be measured.

It is a widely held belief among some people that the uncertainty of inference can somehow be negated by insuring that a "representative" sample is used. Representative sampling, or stratified sampling, is a valuable method of sampling, but its use depends upon having some prior knowledge of the population. Often, however, the term is incorrectly used to refer to sampling when nothing is known about the population. It would be very nice indeed if we knew that the sample being examined "represented" the population of interest, for then everything we observe in the sample would also be valid for the population. The "clinker" in this logic is discovered when one considers the following question, 'How is it determined that the sample represents the population?" That is, if we know that the sample is representative with respect to the attributes we desire to pinpoint, then we already know about them in the population, and taking a sample is a waste of time and effort.

Let us try to determine exactly what is meant, or should be meant by the term " representative sample" for the cases where no prior knowledge of the population is assumed. When one says "representative," he probably has in mind that the sample is "like" the population in the most important respects. That is, by the very use of the term "population," one has implicitly delineated the properties which define the population. He knows whether something is or is not a member of the population by whether or not it possesses those properties. Thus, we see that the question, "Is this a representative sample?" is properly translated, 'Did this sample come from the population we wish to characterize?" This is an important consideration, and is often a factor in erroneous inferences. Generalization from the sample to the population is a risky business at best, but it can be suicidal when the population one actually samples is only a subset of the population he purports to make conclusions about. Thus, we arrive at one of the fundamental rules of inferential statistics, "The legitimate population of inference is that population actually sampled." One should constantly ask himself the question, 'What is the valid population of inference this sample corresponds to?" A reasonable approach to the problem is to first clearly define the population to be studied, and to then make certain that the sampling scheme used allows each member of the population at least a chance to come into the sample.

We should hasten to add, however, that there are times when generalization beyond the legitimate population of inference is apparently unavoidable. We are often forced to come to conclusions or make decisions based on scant data, and we do this by convincing ourselves that the population not sampled is "not very different" from the one sampled. It is a well known fact, for example, that opinion surveys are necessarily restricted to sampling the population of cooperative people who will respond to them, but this does not prevent fairly accurate predictions from being made the majority of the time.

In the case of conceptual populations of repetitions of experiments, the relation between the legitimate population of inference and the experimental design should be clearly understood. To a certain extent, the desired population of inference determines the experimental design, but not completely. However, the experimental design does completely determine the legitimate population of inference. In other words, the design dictates the manner in which the experiment would be repeated, and this, in turn, determines the conceptual population of repetitions. It can be very disconcerting to discover, after an experiment has been conducted, that the population of inference in which the experiment has valid interpretation is far more restricted than originally intended, but it is a frequent consequence of poorly designed experiments. To decrease the likelihood of this happening, one should first get a clear idea of the population to which inferences are to be drawn, and then he should enlist the aid of a competent statistician in designing the experiment so that that population is properly sampled.

We have discussed how one aspect of the problem of getting "good" samples is that of taking the sample from the correct population.

Suppose one has clearly defined the population of interest and realizes that his method of sampling should permit each member of the population a chance to be chosen. (Recall that by "population" and "member," we are including experiments and conceptual repetitions thereof.) What, then, should be the method of taking the sample? The approach, which from the standpoint of valid inference is essential, is one called random sampling. Simply stated, random sampling is the use of a chance mechanism which chooses a sample in such a way that every member of the population has a given probability of being chosen, and a random sample is a sample which has been obtained by such a method. It is a common practice to use simple random sampling, whereby every member of the population is given an equal chance to come into the sample. For experiments and their associated populations of repetitions, obtaining random samples is generally more involved. Here, random sampling is accomplished with a technique referred to in the theory of experimental design as randomization. Any given experimental design has an associated randomization scheme; in fact, the randomization scheme determines the design and thereby the population of inference. Thus, for experimentation, getting a proper sample from the desired population of inference is a matter of proper randomization.

What, then, is the advantage of random sampling? We discussed before the idea of variability of a population, and the notion will be further developed in the next section. For the moment, however, let us agree that, in situations most frequently encountered, different members of the population yield varying measurements and that we may not know exactly what causes some to be high and others low. When the sample is selected systematically according to some criterion, we run the risk of selecting only the high ones or the low ones, for the criterion we use may very well be characteristic of high responses only or of low responses only. Systematic sampling has another drawback also, and it is that the samples drawn in this way often tend to be more homogeneous. This, then, can have the effect of causing the variability in the population of inference to be grossly underestimated which, in turn, leads us to believe our estimates to be more precise than they really are. But, when we use random sampling, chance alone dictates whether or not a particular member comes into the sample. Note that the use of such an approach does not preclude obtaining "unlucky" samples (e.g., all high yields); rather, it allows such events to occur with small probability.

This brings us to a very important property of random sampling or simple random sampling, and that is the long-run nature of the procedure. To put it plainly, there simply is not a technique which will guarantee that a "good" sample will be drawn every time. For the single sample or experiment, random sampling cannot make the results more or less valid; it is only when we place the sample or experiment in the context of being one of a population of conceptual repetitions that random sampling is of any value.

Simple random sampling is applicable for the case when nothing is known about the structure of the population being sampled. Often, however, one does know something about the population and, in such cases, that knowledge can be taken advantage of by the use of restricted random sampling; i.e., sampling which is partly systematic and partly random. For example, we may know from past experience that in measuring some response in a population of people, a certain group tends to give high responses, another group moderate responses, and the remaining group low responses.

Knowing this, we would not want a sample containing only members of any one group, for then we would get a distorted estimate of the response of interest. For such a case, we might require that any sample chosen contain certain proportions of observations from each group. For example, if the sample size is n with n_1, n_2, n_3 to be taken from the three groups, respectively, we may choose n_1 members from the first group at random, n_2 members at random from the second group, and n_3 members at random from the last group. Many similar and more complex situations are possible, and the general rule is that any prior knowledge of response patterns in the population should be taken advantage of by means of restricted random sampling. This is especially true for experiments where restricted randomization is almost always used.

Even though restrictions may be placed on sampling, in order to provide a basis for valid statistical inference, there must remain an element of randomness. The rule is that the sampling procedure must determine for each member of the parent population the probability that it will be chosen in the sample and must be conducted in such a way that population members will be chosen according to that probability. Thus population members may be chosen with equal probability (simple random sampling), or they may be chosen with unequal probabilities (restricted random sampling), but they must be chosen according to some probability; chance must govern the selection process.

3. VARIABILITY AND RANDOM ERROR

We have mentioned variability as a factor one must consider in the study of populations, and it should be emphasized again that one of the major problems that one is faced with in such an endeavor is how to draw reasonable conclusions about the population by observing only a few members, when different members give varying measurements. Let us examine the matter of variability somewhat more closely. To simplify matters, let us focus on conceptual populations of repetitions of experiments. Here we know, for example, that even under conditions as identical as we can make them, two or more experiments do not yield the exact same response. What causes this variation in results? There are at least two ways to view the matter. The deterministic point of view is that if all the circumstances surrounding an event were known, then the result could be predicted with exactitude, or that if time could be turned back and exactly the same set of conditions repeated, the exact same result would be obtained. In other words, the failure of two or more experiments to give the same result is said to be due to an inability to match the set of conditions. The probabilistic approach, on the other hand, is that even if it were possible to obtain identical conditions, the outcome would still be subject to variation because of an underlying chance process. Any distinction between the two philosophies is unimportant to the scientist, however, because from a practical standpoint, they both lead to the same place. Since it is, in fact, humanly impossible to identify all of the sources of variation affecting an event, one simply recognizes and isolates a few of the important sources of variation, calling these the "set of conditions," and the remaining variability is said to be due to random error. Thus, regardless of whether random error is perceived to be an inherent quality in nature, or the result of the combined (and chaotic) effect of undetermined conditions, or even a combination of both of these, the end result is the same. That is, in both cases, we postulate the existence of random error and attribute unexplained variation to its effect. It is immediately seen, however, that unless we are willing to also assume something about the behavior of random error, we can make very little use of the presumption of its existence. We therefore go another step further and assume that the behavior of random error is governed by the laws of probability. This latter assumption is a significant one and is one of the fundamental reasons why statistics has had such an impact upon almost every area of scientific experimentation. It appears that, almost without exception, one can do a better job of describing the physical world with models which have a probabilistic component to them rather than a mathematical (deterministic) component alone, in some cases even to the exclusion of a mathematical component altogether.

Let us now consider for a moment a few of the unfortunate consequences which can sometimes result when people are unwilling to go any further than to grudgingly admit that their experimental results may be subject to "some" uncontrolled variation. A common attitude in this instance is that any acknowledgement of "error" is a reflection upon technique, and suitable refinement of technique will serve to make the error "negligible," There are several potential pitfalls which can be precipitated by such an attitude. One is that the error may not be negligible at all. Recall that, in most cases, experimentation is done in order to try to determine what happens "in general," or what will happen if the experiment is repeated at some future date under the same general conditions, and the error which we should be thinking of is the variation in results which should be expected when the experiment is repeated anywhere within the restrictions which define the population of inference (experimental error). All too often, when someone claims that the error is negligible, he has in mind the variation observed when multiple measurements are taken (sampling error). Almost without exception, the magnitude of the error is smaller in the latter instance. Thus, it is possible for sampling error to be negligible while experimental error is considerable. Another slightly different problem which can arise, whenever one tries to eliminate error by refinement of experimental technique, is that an experiment can often be rendered worthless by over-restriction. In other words, the conditions can be made so specialized so as to preclude generalization to the broader class of conditions of real interest. In addition, it sometimes happens that entirely too much time and money is wasted by trying to reduce the error further than it really needs to be. One of the lessons the statistical approach has taught us is that it is possible to make sense of results in spite of the presence of error, provided it does not cloud the issue too greatly. Thus, one should be concerned with the magnitude of the error in relation to the magnitude of the response he is trying to study, rather than with the absolute magnitude of the error alone.

The most satisfactory way to avoid the above complications, as well as others, seems to be the use of a probability model to represent what we believe to be the behavior of random error. In adopting this approach, it is necessary to make an educated guess as to the probability law (usually formulated in rather general terms) which most nearly fits the situation. There are many alternatives to choose from, and the study of them comprises some of the subject matter of mathematical statistics and distribution theory. Some of the concepts associated with probability theory in general, as well as some commonly encountered distributions, will be discussed later. The question which naturally arises, of course, is, "How does one make a reasonable choice with respect to the probability law which is supposed to apply in a given situation?" It should be obvious that we can never really know for sure whether the model chosen is the correct one, but even so, it is still possible to take a scientific approach to the matter. That is, on the basis of the underlying theory, the data itself, intuition, and good judgement, the probability model is chosen. The choice is never considered irrevocable, however, and a particular model is kept only so long as it is not inconsistent with observations. By way of clarification, we should point out that in the large majority of cases, one is not required to go through the process of choosing a model. Most of the time, past experience by previous researchers has indicated the most appropriate model, and the only responsibility one must bear is to verify that his data and the model are not grossly inconsistent. One will discover that a great many of the statistical methods and techniques employed are based on the assumption of one particular probability model, the normal distribution. There is much justification for assuming a normal distribution for random error.

The point, then, to be derived from the preceding discussion for those who wish to apply statistical methods in experimentation, but do not want to become entangled in the underlying theory, is the following: uncontrolled or unexplained variation is a problem in experimentation which cannot be ignored if useful and meaningful results are among the primary goals of the experiment. The most satisfactory approach to the problem found to date is to attribute such variation to random error and to employ some probability model for a structure within which the effects of random error upon the experimental results can be assessed. Although it is usually unnecessary to become overly involved with the underlying model, it is a recommended practice to verify that required assumptions are not blatantly violated and to make a few quick checks to see that the data itself does not give evidence of non-conformity.

As a final note, in answer to those who might question the correctness or appropriateness of using the abstractions of probability and probability models to study physical phenomena, we must reply that, while such objections are valid, they do not apply only to the present instance but to every case where the behavior of the observable physical world is explained, idealized, and simplified in terms of an abstract model. Thus, in order to pass judgement in this situation, we should use the criteria which have been applied in the past to similar situations, and we must therefore consider the use of probability models as the latest (not necessarily the last) step in an evolutionary process. In that context, it may be said without equivocation that the use of probability and probability models has represented a giant step forward in approaches to data collection, data analysis, and data interpretation, and this, in turn, has had a significant impact in a great many areas of scientific endeavor.

4. BASIC CONCEPTS OF PROBABILITY AND MATHEMATICAL STATISTICS

To understand statistical methods, one should know something of the mathematical concepts behind them. In this section, we shall take a brief look at a few of the basic ideas from mathematical statistics and in the next section, we shall show how they can be applied to some simple problems of inference. For those who are familiar with elementary probability theory and mathematical statistics, this section may be either skipped entirely or used for purposes of quick review.

4.1 Random Experiments and Probability

For the purposes of this section, by a random experiment we shall mean any observable phenomenon, which can be repeatedly observed, with varying results attributable to chance. We call the results of a random experiment its outcomes and the set of all possible outcomes, we call the sample space of the random experiment. When outcomes are combined into sets, we call these combinations of outcomes events. Suppose we are interested in a particular event. If the outcome of the random experiment is contained in this event, then we say that the event has occurred.

Example 1: Let a random experiment consist of tossing a pair of dice, one white and one black. The outcomes have to do with the number showing on the top faces of the dice. Let us denote the outcomes by writing (i,j), where i is the number on the white die and j is the number on the black die. For this random experiment, the sample space consists of 36 possible outcomes, which we may identify as:

$$\begin{array}{lll}
(1,\ 1) & (2,\ 1)\ .\ .\ .\ .\ . & (6,\ 1) \\
(1,\ 2) & (2,\ 2)\ .\ .\ .\ .\ . & (6,\ 2) \\
\ \ \cdot & & \\
\ \ \cdot & & \\
\ \ \cdot & & \\
(1,\ 6) & (2,\ 6)\ .\ .\ .\ .\ . & (6,\ 6)
\end{array}$$

Let A be the event that the sum of i and j is 7. Then, we can write: $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, so A is made up of 6 outcomes, and A occurs whenever a toss of the dice results in any one of the 6 outcomes.

Since a random experiment has varying outcomes, any event (except the event containing all possible outcomes and the event containing no outcomes) will sometimes occur and will sometimes not occur. In order to somehow describe what is to be expected with regard to the occurrence of an event, we use the concept of probability and speak of the probability of occurrence of the event. For the case above, it seems reasonable to think as follows: If the dice are evenly balanced (not "loaded"), then on a single die, any one face should come up as often as another. Since the die has six faces, each face should come up 1/6 of the time. For the pair of dice, the reasoning is similar. There are 36 "faces" possible, each equally likely, so that any one "face" will come up 1/36 of the time. The proportion of the time that the event A, described above, will occur is 6/36 or 1/6, because A contains 6 of the 36 equally likely outcomes.

One should note that the reasoning process used above does not *define* probability; it merely described an *interpretation* of it for the problem at hand. Unfortunately, here, as in many cases, it is possible to define the concept in exact mathematical terms, and yet have no guidance as to how it is to be used or interpreted in practice. In pure mathematics, it is permissible (and necessary) to define a concept, say C, by stating what essential properties it must possess, without going into a philosophical discussion of the type, "What is C?". Thus, while everyone agrees on the mathematical properties that probability must have, there is some disagreement as to how the question, "What is probability?", should be answered. More on that later, but now let us look at some properties of probability.

Before going any further, we must understand events a little better. First, we must recognize that the whole sample space itself can be considered an event; call this event S. We must also allow impossible occurrences to be regarded as events. If the event B contains no outcomes of S, then B will be said to be empty, and we write $B = \emptyset$. Finally, we must say how events may be combined. There are two operations which are needed, and they will be denoted by $\cap$, read "and" or "intersection," and $\cup$, read "or" or "union". Let A and B be two events. Then by the event $A \cap B$, it is meant those outcomes, and only those outcomes, which are common to both A and B. By the event $A \cup B$, it is meant all outcomes in A and all outcomes in B. In addition, there is negation or complementation; for example, the event "A and not B" means those outcomes in A but not in B. The event "not A" is taken to mean "S and not A".

Example 2: For the dice throwing experiment, let A be the event that the sum of the values on the faces is 7, let B be the event that a 1 is on the white die, and let C be the event of getting a double. Then

$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

$B = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}$

$C = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

and

$A \cap B = \{(1, 6)\}$

$A \cup B = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6,1), (1, 1), (1, 2), (1, 3), (1, 4), (1, 5)\}$

$A \cap C = \emptyset$

A and not $B = \{(2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

If the events A and B have no outcomes in common (i.e., $A \cap B = \emptyset$), then A and B are said to be mutually exclusive. Since $P(A \cap B)$ will be zero in this case, we sometimes say that "If A and B are mutually exclusive events then they cannot occur together."

We may now state some of the mathematical properties of probability. In what follows, we shall use $P(A)$ to mean, "the probability that the event A occurs." Let S be the sample space of outcomes of a random experiment, then probability is a function P defined for all the events of S, with the following properties:

1. P ranges in value from 0 to 1 (0 means impossible, 1 means surety).
2. $P(S) = 1$.
3. For any event A, $P(\text{not } A) = 1 - P(A)$.
4. For events $A_1, A_2, A_3, \ldots$, which are pairwise mutually exclusive, $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$

Actually, probability is something almost everyone has a good intuitive feel for, and there is no great problem in bridging the gap between mathematical definition and practical interpretation. The most widely accepted method of interpretation used to arrive at a satisfactory probability assignment, consistent with the properties above, is the frequency interpretation of probability. Using the frequency interpretation, the probability of occurrence of an event is taken to be the proportion of the time that the event occurs. It is now evident that this method of assignment of probabilities was the one used to obtain the probabilities for the dice-throwing experiment.

Some other features of probability should be introduced at this time, one being the concept of independence of events. Let A and B be two events, necessarily from the same sample space. Then A and B are said to be independent if and only if $P(A \cap B) = P(A) \cdot P(B)$. In words, "A and B are independent whenever the probability that A and B occur together is equal to the product of their respective probabilities." As a general rule, we may classify events as independent whenever the occurrence of one event does not affect the occurrence of the other.

Closely associated with the concept of independence is that of conditional probability. Consider the following situation: from an ordinary deck of 52 cards (completely shuffled), what would be the probability that the first card dealt was the queen of hearts? The answer, of course, is 1/52. Now, suppose someone looked at the card and told us that it was indeed a heart. What odds would you now place on the card being the queen of hearts? Of course, the odds would not still be 52 to 1, because 3/4 of the possibilities have been eliminated. One would surely reason that, since there are 13 hearts, the probability of it being the queen is 1/13.

Let A and B be two events. Then the conditional probability

B occurs, given that A has occurred, written $P(B|A)$ is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad .$$

For the cards, let A be the event of getting a heart, and let B be the event of getting the queen of hearts. Then

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/52}{1/4} = 1/13 \quad .$$

The reason that we can characterize independence in terms of occurrence of events not affecting each other can be easily seen. If A and B are independent, then

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B)$$

and

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A) .$$

The concepts of independence and conditionality have far-reaching consequences in statistical theory, and one who is interested in delving further should consult a textbook in mathematical statistics.

4.2 Random Variables and Distributions

In order to put all random experiments on a common footing and to simplify studying them, it is useful to transform the outcomes of random experiments into numbers, and to think of the outcomes as spread along the number line. A function which transforms the outcomes of a random experiment into numerical results is called a random variable. For a given random experiment, there are usually several alternative ways to effect such a transformation, so that several different random variables can be associated with the same random experiment. Generally, however, we have in mind a particular random variable for a given random experiment. Random variables are commonly denoted with capital letters, X, Y, etc.

Example 3: For the dice-throwing experiment, let the random variable X be the sum of the numbers on the top faces. Then the transformation of the outcomes to numbers is:

Outcomes	Value of X	Probability of Value
(1, 1)	2	1/36
(1, 2), (2, 1)	3	2/36
(1, 3), (2, 2), (3, 1)	4	3/36
(1, 4), (2, 3), (3, 2), (4, 1)	5	4/36
.	.	.
.	.	.
.	.	.
(4, 6), (5, 5), (6, 4)	10	3/36
(5, 6), (6, 5)	11	2/36
(6, 6)	12	1/36

In effect, a new sample space is created, where outcomes are described in terms of the random variable X. Thus, we speak of $P(X = a)$, "the probability that X has the value a" or $P(a < X \leq b)$, "the probability that X is greater than a and less than or equal to b."

For any random variable X, once it is specified what values it is possible for X to have and the probabilities that these values occur are given, then the behavior of X is completely determined. In theory, there are two types of random variables, *discrete* and *continuous*, and the type is determined by the nature of the specifications above. A function which performs these specifications for a random variable is called the *probability function*: it is commonly called a *probability mass function* (pmf) when it describes a discrete random variable; and a *probability density function* (pdf) when it describes a continous random variable. To understand the concepts associated with random variables, it is useful to think of the real number line as a rigid one-dimensional mechanical system with mass distributed along its length. If the entire system is considered to have unit mass, then the analogy between mass and probability is complete. The discrete case can be thought of as one where a finite or, at most, a countable number of discrete points in the system have non-zero mass, and the continuous case can be thought of as one where the mass is spread evenly and continuously along the line, the distribution of mass being described by means of a density function.

The dice-throwing example illustrates a discrete random variable in mathematical terms. Let p be a function which is non-zero for at most a countable number of points on the real line such that:

1. $p(x) \geq 0$ for all x.
2. $\sum_{x} p(x) = 1$.

Then p can be considered to be a probability mass function for some random variable X, and any random variable which has such a probability mass function is said to be a *discrete* random variable. If $p(x)$ is the probability mass function for the random variable X, then for every x, $p(x) = P(X=x)$; that is, $p(x)$ is the probability that X takes the value x. In terms of the mechanical system it is seen that $p(x)$ specifies what points are to receive non-zero mass and apportions the mass to those points.

As a general rule, a discrete random variable will result from experiments where a counting process is used.

Example 4: Consider an experiment where there are only two possible outcomes, zero or one, and suppose $P(\text{zero}) = q$, $P(\text{one}) = p$, with $p + q = 1$. Let the experiment be repeated 5 independent times. If we consider a random experiment to consist of 5 repetitions of the simple zero-one experiment, the outcomes can be represented as a string of zeros and ones, 5 digits long. Let X be the number of ones in each outcome of digit strings; then X is a discrete random variable which can have the values 0, 1, 2, 3, 4, 5. The probability mass function for X may be determined as follows: For $x = 0, 1, 2, 3, 4$, or 5, we must count those outcomes having x ones and $5 - x$ zeros, since we can determine that the probability of obtaining any single such outcome is $p^x q^{5-x}$. By the use of permutation and combination theory, it is determined that for any given x, there will be $(5!)/[x!(5-x)!]$ outcomes having exactly x ones and $5 - x$ zeros. For economy of notation, let $C_x^n = (n!)/[x!(n-x)!]$. Then, we have determined that

$$\begin{aligned} p(x) = P(X = x) &= C_x^5 p^x q^{5-x} \qquad \text{for } x = 0, 1, \ldots, 5 \\ &= 0 \qquad \text{for all other } x. \end{aligned}$$

Is $p(x)$ a proper probability mass function? Well, we certainly have $p(x) \geq 0$ for all x. The only question, then, is whether the "system" has unit "mass." But,

$$\begin{aligned}\sum_x p(x) &= \sum_{x=0}^{5} C_x^5 p^x q^{x-5} \\ &= q^5 + 5pq^4 + 10p^2q^3 + 10p^2q^3 + 5p^4q + p^5 \\ &= (q + p)^5 \\ &= 1, \text{ since } p + q = 1\end{aligned}$$

This example illustrates a special and important type of discrete random variable, the binomial random variable. Some other important discrete random variables will be discussed later.

Whenever outcomes of random experiments are measurements, as opposed to counts, it is convenient to idealize the situation by using continuous random variables. Strictly speaking, a continuous random variable must be considered a mathematical approximation, because the set of realized outcomes of any physical measuring process must be a discrete set. Nevertheless, it is useful to suppose that there are random experiments which yield a continuum of outcomes, if they could only be measured as such. In mathematical terms: Let f be a function defined for all real x, which is continuous except for at most a countable number of points such that

1. $f(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

Then $f(x)$ can be considered to be a probability density function for some random variable X, and any random variable which has such a probability density function is said to be a continuous random variable. If $f(x)$ is the probability density function for some random variable X, then for any a and b with $a < b$, $P(a < x < b) = \int_a^b f(x)dx$. In terms of the mechanical system, a probability density function is completely analogous to a mass density function, and the mass of any section (a, b) is determined as the area under the density curve from a to b.'

Example 5: Consider a process for which occurrences of some event are anticipated, and assume that frequency of occurrence is governed by the following: (1) there is some positive number r such that if a short enough interval is taken, say h, the probability of exactly one occurrence in the interval h is rh; (2) the probability of more than one occurrence in the interval h is practically 0; and (3) the occurrence in any one interval of length h does not affect occurrence in some other non-overlapping interval of length h. Such a process is called a Poisson process with parameter r, and is highly useful in applications involving queueing theory (customer arrivals in a service line, incoming telephone calls to a central operator, flaws along a length of wire, failures in transistors, etc.). It applies equally well to time intervals and to distance intervals. Observing such a process can give rise to both discrete and continuous random variables, depending upon what aspect of the process is observed. Assume that the interval measure is time.

(a) A Discrete Random Variable

Suppose a Poisson process with parameter r is observed for a length of time t, let X be the number of occurrences observed during that time, and let $\lambda = rt$. Then it can be shown that X is a discrete random variable with probability mass function:

$$p(x) = P(X = x) = \frac{(\lambda)^x}{x!} e^{-\lambda} \qquad \text{for } x = 0, 1, 2, \ldots$$

$$= 0 \qquad \text{otherwise.}$$

It is easily determined that $p(x)$ is a proper probability mass function, since

$$\sum_{x=0}^{\infty} \frac{(\lambda)^x}{x!} e^{-\lambda} = e^{-\lambda} (1 + \lambda + (\lambda)^2/2 + (\lambda)^3/3! + \ldots)$$

$$= e^{-\lambda} \cdot e^{\lambda} = 1.$$

A discrete random variable having a probability mass function like p, given above, is called a Poisson random variable with parameter λ.

(b) A Continuous Random Variable

Suppose a Poisson process with parameter r is observed until the first occurrence is noted, and let T be the time it takes for this to happen. Then it can be shown that T is a continuous random variable with probability density function:

$$f(t) = re^{-rt} \qquad \text{for } t > 0$$

$$= 0 \qquad \text{otherwise.}$$

Here, to convince ourselves that f is a valid probability density function, we must verify that the area under the curve of f is 1, and this is easily done, since,

$$\int_0^{\infty} re^{-rt}dt = \int_0^{\infty} e^{-u}du = 1.$$

A continuous random variable having a probability density function like f, given above, is called an exponential random variable with paramenter r.

Although random variables are completely determined by their probability functions, it is sometimes useful to consider another function which, from a mathematical viewpoint, is even more elementary than a probability function. Let X be a random variable (discrete or continuous). The cumulative distribution function of X is defined as:

$$F(x) = P(X \leq x).$$

For the discrete case

$$F(x) = \sum_{t \leq x} p(t) ,$$

where $p(t)$ is the pmf for X, and for the continuous case,

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

where $f(t)$ is the pdf for X.

Random variables are also completely determined by their cumulative distribution function (cdf), and there are times when it is easier to use it. Some properties of a cdf which can be immediately seen are:

1. $0 \leq F(x) \leq 1.$
2. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1.$
3. $F(x)$ is nondecreasing.

The cdf of a discrete random variable is characteristically a step function with jumps at the points having positive probability mass, the size of each jump being equal to the probability mass at that point. The cdf of a continuous random variable, on the other hand, is characteristically a smooth, strictly increasing, lazy-s-shaped curve. Thus, given a cdf, we may determine the corresponding probability functions by $p(x) = F(x) - \lim_{z\to x^-} F(z)$ (the jump of F at x) if discrete, or by

$f(x) = F'(x)$ if continuous. Whenever a random variable X can be specified with some particular probability function or cumulative distribution function, we speak of the *distribution* of X, referring, literally, to how its probability mass is distributed along the real line. Thus, we say that "X has a Poisson distribution," or "X has an exponential distribution," and so on.

4.3 Moments and Expectation

It is natural to try to find ways of describing distributions without having to resort to a full description given by a pmf, pdf, or cdf. When a probability distribution is visualized as a mechanical system, it readily occurs to one to use moments of the system to accomplish this. Consider the first two moments. When a system has unit mass, finding the first moment simply locates the center of mass, which is referred to as the *mean* of the distribution. Thus, for a discrete distribution with pmf $p(x)$,

$$\text{mean} = m = \sum_x xp(x) \quad \text{(sum of positive and negative moments)},$$

and for a continuous distribution with pdf $f(x)$,

$$\text{mean} = m = \int_{-\infty}^{\infty} xf(x)dx.$$

Next, we can compute the second moment about m, which is the moment of inertia of the system, and which we refer to as the *variance* of the distribution. Accordingly, for the discrete case,

$$\text{variance} = \sigma^2 = \sum_x (x - m)^2 p(x),$$

and for the continuous case,

$$\text{variance} = \sigma^2 = \int_{-\infty}^{\infty} (x - m)^2 f(x)dx.$$

For most distributions of interest, these moments exist and, for many distributions, the mean and variance give a reasonably good description; in a few cases, they give a complete description. The mean, of course, gives us a measure of where the distribution is located in terms of its mass, and the variance gives us a measure of the spread of the distribution, in the sense that a small variance indicates that the probability mass tends to be concentrated at the mean, while a larger variance indicates that the probability mass tends to be less concentrated at the mean and more "spread out." There are many instances where it is more convenient to use the square root of the variance. The square root of the variance is called the standard deviation and is denoted by σ.

Consider, now, a distribution which has center of mass at m, but has, say, 90% of its probability mass close to m on the left and 10% spread far out to the right of m. Here, mean and variance do not give a reasonable description, due to the fact that the distribution is skew (not symmetric). For some distributions, more than the first two moments are needed for descriptive purposes, and a combination of second and third moments can be used to describe this lack of symmetry or skewness. For discrete distributions,

$$\text{skewness} = \alpha_3 = \sum_{x} \frac{(x - m)^3 p(x)}{\sigma^3} \quad ,$$

and for continuous distributions,

$$\text{skewness} = \alpha_3 = \int_{-\infty}^{\infty} \frac{(x - m)^3 f(x)dz}{\sigma^3} \quad .$$

For symmetric distributions, $\alpha_3 = 0$, and for others, α_3 is positive or negative depending upon whether the distribution is skewed to the right or skewed to the left ("skewed to the right(left)" means that the bulk of the probability mass is to the left(right) of m).

The calculation of moments is a special case of a more general type of operation called expectation or taking expected value. Let X be a random variable (discrete or continuous), and let $g(x)$ be some function defined over the points in the range of x. Then, we define the expected value of $g(X)$, written $E[g(X)]$, as:

$$E[g(X)] = \sum_{X} g(x)p(x) \quad \text{if } X \text{ is a discrete r.v. with pmf } p(x),$$

or

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dz \quad \text{if } X \text{ is a continuous r.v. with pdf } f(x).$$

Thus, it is immediately seen that by taking $g(x) = x$,

$$\text{mean} = m = E(X),$$

and by taking $g(x) = (x - m)^2$,

$$\text{variance} = \sigma^2 = E[(X - m)^2].$$

Expected value is a mathematical operator, and is highly useful when applied as such. From properties of summation and integration, we may derive the following rules governing expected value:

1. If c is a constant (not a random variable), $E(c) = c$.
2. If c is a constant and X is a random variable, $E(cX) = cE(X)$.
3. If X and Y are both random variables, $E(X + Y) = E(X) + E(Y)$.

By using E as an operator, we may derive, for example,

$$\text{variance} = E[(X - m)^2] = E(X^2 - 2Xm + m^2) = E(X^2) - 2mE(X) + m^2$$
$$= E(X^2) - m^2$$
$$= E(X^2) - E^2(X)$$

Example 6: Consider the exponential distribution, where

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad x > 0$$
$$= 0 \quad \text{otherwise}$$

$$\text{mean} = E(X) = \int_0^\infty xe^{-\lambda x}dx = 1/\lambda \int_0^\infty (\lambda x)e^{-(\lambda x)}d(\lambda x) = 1/\lambda \int_0^\infty ue^{-u}du = 1/\lambda$$

$$\text{variance} = E[(X - m)^2] = E(X^2) - m^2,$$

$$\text{but,} \quad E(X^2) = \int_0^\infty x^2e^{-\lambda x}dx = 1/\lambda^2 \int_0^\infty (\lambda x)^2e^{-\lambda x}d(\lambda x) = 1/\lambda^2 \int_0^\infty u^2e^{-u}du = 2/\lambda^2$$

so, variance $= 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$.

Since variance is defined in terms of expectation, we may also consider variance as an operator, and it is useful to do so. Let us denote it by Var ( ) and note that the following rules can be derived:

1. If c is a constant, (not a random variable), $\text{Var}(c) = 0$.
2. If c is a constant and X is a random variable, $\text{Var}(cX) = c^2\text{Var}(X)$.

4.4 Other Descriptive Quantities

In addition to mean, variance, and skewness, there are other attributes of distributions which can aid in describing them. Two such quantities are the median and the mode. For some distributions, these quantities may or may not be unique, but no essential difficulties arise as a result of this fact. A median is any point on the x-axis which divides the distribution in half; that is, 50% of the probability mass lies on either side of the median. For a discrete distribution, working from the lower end, it frequently happens that a point x will be reached where including p(x) puts more than half of the distribution on the left end, while leaving p(x) out puts more than half of the distribution on the right. In such a case, x or any point close by will do fine for a median. For a continuous distribution the median will always be unique. We might note, in passing, that the position of the median relative to the mean can be used as an indication of non-symmetry; for example, if the median is to the left of the mean, the distribution is skewed to the left.

As a generalization of the idea of a median, one may consider the use of *distribution quantiles*. For example, the quantiles divide the distribution into fourths, the deciles into tenths, the percentiles into hundredths, and so forth. Distribution quantiles are interesting in that they give measures of both location and spread of a distribution. In addition, their behavior can be mathematically characterized independently of what distribution they arise from, and for that reason, techniques based on properties of distribution quantiles are said to be *distribution-free*.

The use of *modes* to describe distributions is less common, but they will be included here for the sake of completeness. For a continuous distribution, the mode or modes are simply the points at which the relative maxima of the probability density function occur. Most continuous distributions of practical interest have a single mode. Most theoretical discrete distributions also have a single mode, and in these cases the mode is the value of x for which p(x) is maximum. If a distribution has a single mode, it is said to be unimodal, and if it has two (or more) modes, it is said to be bimodal (multimodal).

4.5 Jointly Distributed Random Variables

Sometimes the outcomes of interest of random experiments cannot be adequately represented by single numbers, but require sets of numbers. Suppose, for the moment, that the outcomes of some random experiment can be represented by a pair (x, y). To handle this situation, we use the concept of a two-dimensional random variable, say (X, Y), where each of the components is a one-dimensional random variable. In the sample space of (X, Y), events are of the type $(a < X < b,\ c < Y < d)$.

Example 7: In the dice-tossing experiment of Example 1, let X be the number showing on the white die, and let Y be the number showing on the black die, and suppose we are interested in what comes up on both dice. The values of (X, Y) which can occur are, of course, the 36 pairs (x, y) with $x = 1, 2, \ldots, 6$ and $y = 1, 2, \ldots, 6$, and each of the (x, y) pairs occurs with the same probability. For this experiment, we would say that X and Y are discrete random variables with joint probability mass function

$$\begin{aligned} p(x, y) = P(X = x \text{ and } Y = y) &= 1/36 \quad \text{for } x = 1, 2, \ldots, 6 \text{ and} \\ &\phantom{= 1/36 \quad \text{for }} y = 1, 2, \ldots, 6 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The example illustrates a joint distribution of discrete random variables. Let us now restrict ourselves to jointly distributed continuous random variables. Distributions of this type may be visualized as surfaces over the xy-plane with the height of the surface at the point (x, y) representing the density of probability at that point. The joint probability density functions of random variables must conform to the same rules as before; that is, they must satisfy:

1. $f(x,y) \geq 0$ for all x and y.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$

The joint pdf of X and Y is used to determine probabilities of events in a manner analogous to the one-dimensional case. For example,

$$P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y)dydx$$

or

$$P(a < X < b) = \int_a^b \int_{-\infty}^{\infty} f(x,y)dydx$$

or

$$P(Y < d) = \int_{-\infty}^{d} \int_{-\infty}^{\infty} f(x,y)dxdy.$$

We may also define the joint cumulative distribution function of X and Y as:

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dvdu.$$

Whenever random variables are distributed jointly, it is frequently of interest to know how they are distributed individually. To describe how X and Y are distributed individually, we use the concept of marginal distributions. Suppose the joint distribution of X and Y is known, and we want the marginal distribution of X. Consider some fixed value of x. To determine the value of the marginal probability density function of X at x_0, we "consolidate" the probability along the line $x = x_0$ by integrating over y. Doing this for all x, we would obtain

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

for the marginal pdf of X. By similar reasoning, we would also have

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

for the marginal pdf of Y. The properties of multiple integration guarantee that f_1 and f_2 satisfy the conditions for the probability density functions.

An interesting question now arises: If we know how X and Y are distributed marginally, do we know how they are distributed jointly? The answer is sometimes, but not always, and it depends upon whether or not X and Y are independent random variables. The random variables X and Y are said to be independent if and only if

$$P(a < X < b \text{ and } c < Y < d) = P(a < X < b) \cdot P(c < Y < d).$$

Note that independence of random variables is defined in terms of independence of events; let the events A and B be defined as: $A = (a < X < b)$ and $B = (c < Y < d)$. Then the condition of independence may be stated: $P(A \cap B) = P(A) \cdot P(B)$.

For continuous random variables X and Y having joint pdf $f(x,y)$ and marginal pdf's $f_1(x)$ and $f_2(y)$, it can be mathematically proven that X and Y are independent if

$$f(x,y) = f_1(x)f_2(y) \qquad \text{for all } x \text{ and } y$$

Thus, if X and Y are independent, knowing their marginal distributions is sufficient to determine their joint distribution.

Descriptions of joint distribtuions, for the most part, are analogous to the one-dimensional case. Thus we use

$$\mu_x = E(X) = \int_{-\infty}^{\infty} x f_1(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)dydx$$

$$\mu_y = E(Y) = \int_{-\infty}^{\infty} y f_2(y)dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)dydx$$

$$\sigma_x^2 = Var(X) = E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} (x-\mu_x)^2 f_1(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)^2 f(x,y)dydx$$

$$\sigma_y^2 = Var(Y) = E[(Y - \mu_y)^2] = \int_{-\infty}^{\infty} (y-\mu_y)^2 f_2(y)dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-\mu_y)^2 f(x,y)dxdy.$$

Whenever jointly distributed random variables are not independent, the marginal descriptions are not enough, however, and the concept of covariance is needed. The covariance of the jointly distributed random variables X and Y, denoted by Cov(X, Y) and σ_{xy}, is defined as:

$$\sigma_{xy} = \text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)\,dxdy$$

The covariance of X and Y gives us a measure of how X and Y vary together. The correlation of X and Y, denoted by ρ, is a "standardized" covariance,

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} .$$

Dividing by $\sigma_x \cdot \sigma_y$ has the effect of reducing σ_{xy} to "standardized units" so that, for any joint distribution, $-1 \leq \rho \leq 1$, regardless of the form of the distribution.

The concepts of independence and correlation are closely associated with each other, and are frequently confused. Whenever random variables are independent, $\rho = 0$, but the converse is not always true. Thus, uncorrelated random variables are not necessarily independent random variables.

The development of this section was done for the case of continuous random variables, but the development for discrete distributions is completely analogous. Also, it is possible to consider more than two jointly distributed random variables, with a similar development. The mathematics of n-dimensional random variables is no more difficult than that of two-dimensional ones; however, the geometrical visualizations become almost impossible.

4.6 Useful Distributions

There are some random variables whose distributions are encountered repeatedly in applications, and in this section, some of them will be discussed briefly.

4.6.1. The Binomial Distribution

A Bernoulli trial is a random experiment which has only two possible outcomes, generally called success and failure. Let X be the number of successes in n independent repetitions of a Bernoulli trial, where P(success) = p, for each trial. Then X is a discrete random variable having binomial distribution with parameters n and p. The probability mass function for X is

$$p(x) = \frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

$$= 0 \quad \text{otherwise.}$$

It can be verified that the mean and variance are,

$$\mu_X = E(X) = np$$

$$\sigma_X^2 = E[(X - np)^2] = np(1-p).$$

The cumulative distribution function for the binomial distribution is very tedious to calculate, because sums of terms of a binomial expansion are involved, and for that reason, tables of the binomial cdf have been computed for various values of n and p. Whenever one must do calculations for the binomial distribution, such a table is generally consulted.

Example 8: Let a "fair" coin be tossed 8 times in succession, and let X be the number of heads which would be obtained in those 8 tosses. Then X has a binomial distribution with parameters n = 8 and p = 1/2. The probability mass function for X would be

$$\begin{aligned} p(x) &= \frac{8!}{x!(8 - x)!}\,(1/2)^8 \qquad \text{for } x = 0, 1, 2, \ldots, 8 \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned}$$

and the distribution for X could be tabulated as,

value of X	0	1	2	3	4	5	6	7	8
P(X=x)=p(x)	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$

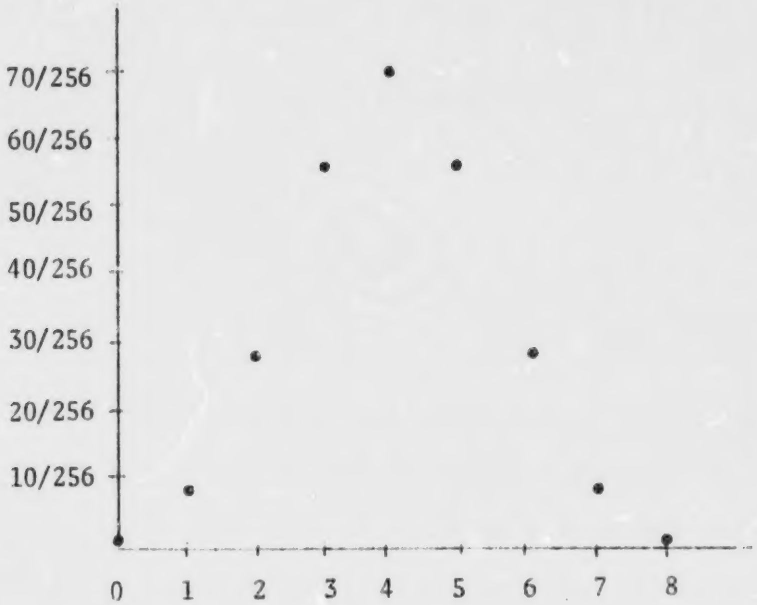


Figure 1. Binomial probability mass function.

A binomial distribution is symmetric about the mean np if $p = 1/2$, and is skewed for other values of p.

4.6.2 The Poisson Distribution

As shown earlier in Example 5, a Poisson random variable can be obtained by observing a Poisson process with parameter r. Suppose the process is repeatedly observed for a time t each time, and let X be the number of occurrences in each time interval of length t. Then X is a discrete random variable and has a Poisson distribution with parameter $\lambda = rt$. The probability mass function for X is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \text{for } x = 0, 1, 2, \ldots$$

$$= 0 \qquad \text{otherwise.}$$

We can compute,

$$\mu_X = E(X) = \lambda$$

$$\sigma_X^2 = E[(X - \lambda t)^2] = \lambda .$$

With $E(X) = \lambda$, we can give a physical interpretation to r. Since the average or expected number of occurrences in time t is $\lambda = rt$, the expected number of occurrences per unit time is r. Thus, we may refer to r as the occurrence rate of the process.

The cumulative distribution function of the Poisson distribution has also been tabulated in standard tables for various values of λ.

Example 9: The occurrence of failures in large lots of transistors can often be represented by a Poisson process. Suppose a particular lot has a failure rate of 1/2000 per hour (one every 2000 hours), and suppose the lot is to be tested for 1000 hours. What is the probability that no more than 2 failures will occur?

If we let X be the number of failures in 1000 hours, then X has a Poisson distribution with parameter $\lambda = 1000\cdot(1/2000) = 1/2$. The probability mass function is

$$p(x) = \frac{1}{2^x x!} e^{-1/2} \qquad \text{for } x = 0, 1, 2, \ldots$$
$$= 0 \qquad \text{otherwise.}$$

The probability sought is $P(X \le 2)$. But

$$\begin{aligned} P(X \le 2) = P(X = 0 \text{ or } X = 1 \text{ or } X = 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= p(0) + p(1) + p(2) \\ &= e^{-1/2}\,(1 + 1/2 + 1/6) \\ &= 0.9856. \end{aligned}$$

Note that if we had desired, we could have consulted a table of the Poisson cdf, $F(x)$, for $\lambda = 1/2$ and $x = 2$.

4.6.3 The Geometric Distribution.

Let a Bernoulli trial, with P(success) = p, be independently repeated until the first success is obtained, and let X be the number of trials necessary to do this. Then X is a discrete random variable having a geometric distribution with parameter p. The probability mass function for X is

$$p(x) = p(1 - p)^{x-1} \qquad \text{for } x = 0, 1, 2, \ldots$$
$$= 0 \qquad \text{otherwise}$$

and

$$\mu_X = E(X) = 1/p$$
$$\sigma_X^2 = E[(X - 1/p)^2] = (1 - p)/p^2 .$$

The pmf of the geometric distribution is actually very easy to derive. If x trials are required to get the first success, then the first $x - 1$ trials must be failures and the xth trial a success. The probability of failures in every one of the first $x - 1$ trials is $(1-p)(1-p)\ldots(1-p) = (1-p)^{x-1}$ and the probability of success on the xth trial is p. Thus, $P(\text{taking } x \text{ trials}) = P(X=x) = p(1-p)^{x-1}$. The geometric distribution is a special case of a more general type of distribution known as the negative binomial distribution.

4.6.4 The Exponential Distribution

This distribution has also been discussed earlier as an example of a continuous distribution (Example 5). If a Poisson process with occurrence rate r is observed until the first occurrence, and if X is the elapsed time until the first occurrence, then X is a continuous random variable having the exponential distribution with parameter r. The probability density function for X is,

$$\begin{aligned} f(x) &= re^{-rx} && \text{for } x > 0 \\ &= 0 && \text{otherwise,} \end{aligned}$$

and

$$\mu_X = E(X) = 1/r$$

$$\sigma_X^2 = E[(X - 1/r)^2] = 1/r^2.$$

One might have suspected that the average time to the first occurrence $E(X)$ would be $1/r$, since the occurrence rate is r per unit time. The similarity of the discrete geometric distribution and the continuous exponential distribution should be noted. Like the geometric distribution, the exponential distribution is a special case of a more general family of distributions called the gamma distributions. The gamma distribution is a two-parameter distribution with probability density function,

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-x/\beta} \qquad \text{for } x < 0$$

$$= 0 \qquad \text{otherwise,}$$

where $\alpha, \beta > 0$. The exponential distribution is obtained by taking $\alpha = 1$ and $\beta = 1/r$.

4.6.5 The Normal Distribution.

The normal distribution is by far the most familiar and most used distribution in applications. Let X be a random variable having the normal distribution. Then the probability density function for X is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-u)^2/2\sigma^2} \qquad \text{for all real } x,$$

where $\sigma^2 > 0$ and $-\infty < \mu < \infty$. We get

$$\mu_X = E(X) = \mu$$

$$\sigma_X^2 = E[(X-\mu)^2] = \sigma^2 .$$

The normal pdf is the familiar "bell curve" shown in Figure 2. It is symmetric about the point $x = \mu$ and has its maximum value there. The curve has two points of inflection at $\mu - \sigma$ and $\mu + \sigma$, and is more sharply peaked for smaller values of σ^2 and flatter for larger values of σ^2.

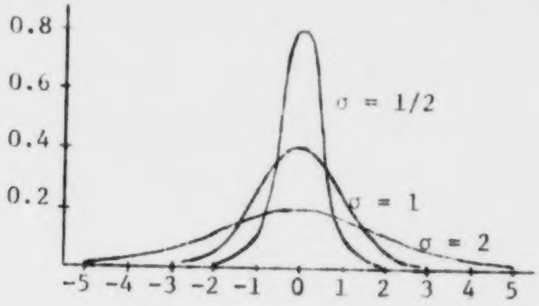


Figure 2. The normal density curve.

If $\mu = 0$ and $\sigma^2 = 1$, we have

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and any random variable having such a pdf is said to have the standard normal distribution. The cumulative distribution function for a standard normal distribution is

$$F(x) = P(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt,$$

which is recognized as a form of the well-known erf function. A closed-form functional expression for F does not exist, and tables of the values of F have been computed, but only for the standard case of $\mu = 0$ and $\sigma^2 = 1$. The reason why further tables are unnecessary lies in the following result. Let X be normally distributed with mean μ and variance σ^2. (For economy, this can be written as, $X \sim N(\mu, \sigma^2)$.) Let $Z = (X-\mu)/\sigma$, then Z is also a random variable and $Z \sim N(0, 1)$. In other words, subtracting the mean μ, and dividing by the standard deviation σ "standardizes" any normal (μ,σ^2) random variable to a normal (0,1) random variable. This result can be used in the following way. Suppose $X \sim N(\mu,\sigma^2)$ and we want to know $F_X(c)$. We can compute

$$\begin{aligned} F_X(c) = P(X < c) = P(X - \mu < c - \mu) &= P[(X-\mu)/\sigma < (c-\mu)/\sigma] \\ &= P[Z < (c-\mu)/\sigma] \\ &= F_Z[(c-\mu)/\sigma], \end{aligned}$$

and since $Z \sim N(0,1)$, F_Z can be extracted from a table of the standard normal cumulative distribution.

Because of the relation between X and Z above, it is a common practice to think of values of a non-standard normal variable in terms of σ-units deviation from the mean μ. For $Z \sim N(0, 1)$, $P(-1 < Z < 1) = 0.683$; thus we can say that for any normal distribution, 68.3% of the observations lie between $\mu - \sigma$ and $\mu + \sigma$ and, in a similar manner, we can determine that about 95.5% of the observations lie between $\mu - 2\sigma$ and $\mu + 2\sigma$ and that 99.6% of the observations are between $\mu - 3\sigma$ and $\mu + 3\sigma$.

One of the reasons that the normal distribution is so useful is the fact that it can be used for accurate approximation of other probability distributions under certain conditions. For example, if X is a binomial (n,p) random variable, then the random variable $(X - np)/\sqrt{np(1-p)}$ is approximately normally distributed for large values of n, and the approximation improves with increasing n. The approximation is also better for p close to $1/2$, since the binomial distribution is more symmetric for p near $1/2$. The process of subtracting np and dividing by $\sqrt{np(1-p)}$ can be thought of as standardization, since $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$.

Example 10: Let X have a binomial distribution with $n = 8$ and $p = 1/2$. We will use the standard normal distribution to approximate $P(X \leq 6)$. In applying the approximation, one step in the technique is that of making a "correction for continuity." To understand why this improves the approximation, it is useful to use a histogram to represent the pmf of the binomial. A histogram is a graphical device used to represent discrete distributions for which the pmf is positive at evenly spaced points along the number line (such as at the integers, 0, 1, 2, ...). Rectangles of constant width are drawn above and centered over the points of positive probability mass in such a way that the areas of the rectangles are proportional to the probability masses of the points. A histogram is given below in Figure 3 for the binomial with $n = 8$, $p = 1/2$. Also drawn is a smooth curve, which we will suppose is an approximating curve.

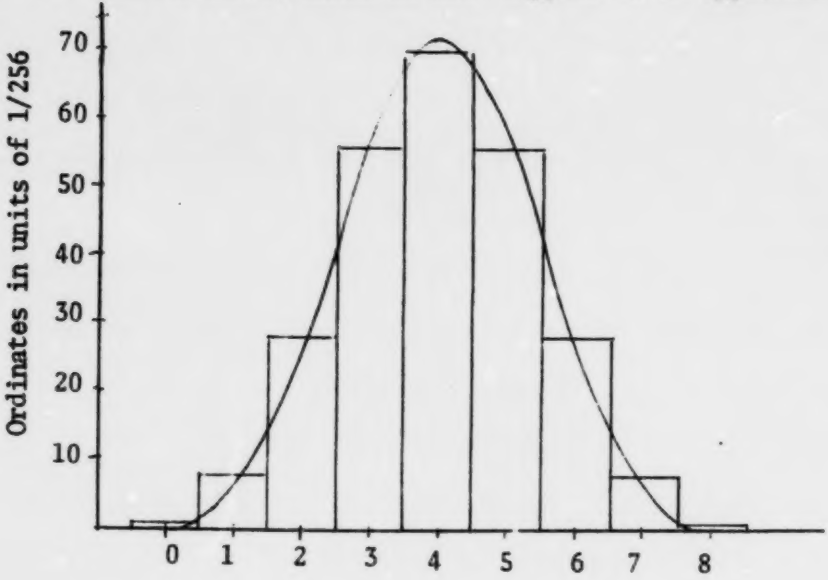


Figure 3. Binomial mass function.

Now, to find $P(X \leq 6)$ from the histogram, we would add the areas of the 7 rectangles at 0, 1, 2, ..., 6 together. To approximate this area by means of the curve, one would compute the area under it to the left of 6.5. Let us suppose that the curve is the graph of the pdf of a random variable Y which is approximately normally distributed with mean $np = 4$ and standard deviation $np(1-p) = 2$. Then

$$
\begin{aligned}
P(X \leq 6) \doteq P(Y < 6.5) &= P[(Y-4)/\sqrt{2} < (6.5-4.0)/\sqrt{2}] \\
&= P(Z < 1.77), \text{ where } Z \sim N(0,1) \\
&= 0.9616.
\end{aligned}
$$

The actual probability obtained from the histogram is

$$P(X \leq 6) = 247/256 = 0.9648.$$

The procedure of finding the area under the curve to the left of 6.5 instead of 6.0 is an illustration of the use of the correction for continuity. The matter of a continuity correction is often a source of confusion to people, but a graph such as Figure 3 illustrates that it can be visualized simply as a matter of including (or excluding) whole rectangles, which implies the use of boundary points, rather than midpoints along the x axis. For example,

$$P(X \leq 6) \doteq P(Y < 5.5) \quad \text{and} \quad P(2 < X \leq 5) = P(2.5 < X < 5.5).$$

A reason often given for using the normal distribution is the convenience of doing so. The mathematical properties of the normal distribution are very "nice," indeed, and a large body of known mathematical facts has consequently been built up around it. While convenience and tractability are important considerations, they are not the only influencing factors in the matter. One of the most important and intriguing theorems in all of statistical theory is the Central Limit Theorem. Roughly stated, the theorem goes as follows: let $X_1, X_2, \ldots, X_n$ be random variables all having the same distribution with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$; and let $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$. Then the random variable $\bar{X}$ has an approximate normal distribution with mean μ and variance σ^2/n. The remarkable thing about the theorem is that the distribution for the X_i's is not restricted to any particular form; any discrete or continuous distribution is allowed. In short, part of the reason for the widespread use of the normal distribution in applications is the Central Limit Theorem.

4.6.6. The Chi-Square Distribution

Let $X_1, X_2, \ldots, X_k$ be independent random variables, each normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 1$, and let $X = X_1^2 + X_2^2 + \ldots + X_k^2$. Then X is a continuous random variable having the Chi-Square distribution with parameter k, and we write $X \sim \chi^2(k)$. The parameter in the Chi-Square distribution is called the degrees of freedom, for reasons which must be explained later. If $X \sim \chi^2(k)$, the probability density function for X is,

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{k/2-1} e^{-x/2} \qquad \text{for } x > 0$$

$$= 0 \qquad \text{elsewhere}$$

with

$$\mu_x = E(X) = k$$

and

$$\sigma_x^2 = E[(X-k)^2] = 2k.$$

Note that the Chi-Square distribution is also a special case of the gamma distribution with $\alpha = k/2$, $\beta = 2$.

4.6.7. Student's t-Distribution

Let $X \sim N(0,1)$ independently of $Y \sim \chi^2(r)$, and let $T = S/\sqrt{Y/r}$. Then T has a Student's t-distribution with r degrees of freedom, and we write T t(r). The actual form of the probability density function of the t-distribution is of secondary interest, and will not be given here. The t-distribution has its mean at zero, is symmetric about that point, and can be visualized as a "spread-out" standard normal curve.

4.6.8. The F-Distribution

Let $X \sim X^2(f_1)$ independently of $Y \sim X^2(f_2)$, and let $F = f_2X/f_1Y$. Then F has an F-distribution with f_1 and f_2 degrees of freedom, and we write $F \sim F(f_1, f_2)$. The mathematical form of the probability density function of the F-distribution is also of secondary interest, and will likewise be omitted.

The short discussion of the Chi-Square, Student's t, and F distributions should not be taken to imply that they are not important distributions, because the opposite is the case. These distributions are called sampling distributions because they were discovered and derived as a result of studying the behavior of sample statistics. Our real interest in these last three types of random variables is in the nature of their cumulative distribution functions, which have been extensively tabulated. A t-table, an F-table, and a Chi-Square-table are standard in almost every text of applied statistics.

5. SAMPLING AND INFERENCE

In this section, we shall examine some of the elementary activities commonly associated with statistical inferences and try to show how probability theory is used to assist in these activities. When we speak of inference, or an inference situation, we will be referring to a situation where one desires to make statements or guesses about a population by examining a subset of it which is believed to somehow typify or represent the population as a whole. Some of the difficulties involved with the terms "typical" and "representative" have been mentioned already and we have seen that about all one can do is to try to insure that a random sample is drawn from the population of inference. Beyond that, "typicality" is largely a matter of faith.

5.1 Description of Finite Number Populations

The function of condensing and summarizing the information in a set of numbers is one which can be traced to the early beginnings of statistics as a discipline. If interest is only in the set of numbers itself, it is quite evident that no statistical inference is involved, since generalization beyond the number population is not required. However, some of the same activities used in describing finite number populations have also been found to be useful in inference situations, and for that reason, we shall mention some of the informative operations one can perform upon a set of numbers. Let the numbers be denoted by $x_1, x_2, \ldots, x_N$; then one may:

1. Rank the numbers in order of increasing magnitude.
2. Compute the mean $m = \sum_{1}^{N} x_i/N$.
3. Compute the range $= x_{max} - x_{min}$.
4. Compute the variance of $\sigma^2 = \frac{\Sigma(x_i-\bar{x})^2}{N}$ or the standard deviation σ.
5. Find the median, quartiles, deciles, etc.
6. Select classes, construct a frequency table, and draw a histogram.
7. Calculate a cumulative frequency function.

It should be emphasized that each of the activities mentioned above retains only a part of the full information contained in the number population. This being the case, it is not difficult to see how a number population could be misrepresented by a "judicious" choice of descriptive techniques.

5.2 Statistical Inference in Finite Populations

In a very real sense, all populations generated by measurement are finite populations, due to the physical limitation of measuring instruments, so the distinction being made here between finite and infinite populations is whether or not a population to be studied is considered to be a finite subset of a [illegible] infinite population of similar members.

For the [illegible] assume that one desires to describe a large population but does not wish to view it as infinite, and assume that it is not possible in terms of time, money, or both to examine the entire population. The experimenter is, of course, confronted with an inference situation, for he must form his opinions about the population by examining one or more samples from it. To simplify matters, let us suppose that the goal is estimation of the population mean μ . It is decided that when the sample $(x_1, x_2, \ldots, x_n)$ is obtained, the estimate of μ will be the sample mean $\bar{x} = \sum_1^n x_i/n$. Is this a good thing to do? We have seen that it is not possible to judge on the basis of the sample itself, because $\bar{x}$ might be "right on the mark" if the sample values $(x_1, x_2, \ldots, x_n)$ happened to be evenly situated close to μ , or $\bar{x}$ could be "badly off" if the sample values happened to be much higher or lower than μ . Moreover, the fact that the value of μ is not known precludes knowing what the actual situation might be. One must therefore resort to an evaluation of the long-run consequences of his procedure; that is, if he continued to take samples and compute $\bar{x}$'s, how would his $\bar{x}$'s, taken as a whole group, measure up? There are at least two things to look for. First, would the $\bar{x}$'s tend to cluster in such a way so as to not consistently overestimate or underestimate μ , and second, would the variation in the $\bar{x}$'s from sample to sample be large or small? In short, what would be the *accuracy* and the *precision* of the estimation procedure? It is not possible to say anything definite about either of these properties, unless the method of taking the samples can be given a probability structure; that is, unless random sampling is required. Let us consider the matter of accuracy; over all the possible samples of size n , what do the associated $\bar{x}$'s average out to?

A quantity such as $\bar{x}$, when regarded as an estimate of a population parameter, say θ , is often denoted by $\hat{\theta}$ whenever the particular estimate is not of immediate interest, and we say that $\hat{\theta}$ is θ, in other words, if $E(\hat{\theta}) = \theta$; otherwise, $\hat{\theta}$ is said to be a biased estimate with Bias = $E(\hat{\theta}) - \theta$. Thus, bias is simply another way of describing accuracy.

Recall that simple random sampling is taking random samples in such a way that each member of the population has the same chance of being included in the sample. One of the primary reasons for using simple random sampling is that the result of doing so makes $\bar{x}$ an unbiased estimate of μ; the average of $\bar{x}$'s over all possible samples of size n is μ. Furthermore, the validity of this assertion does not depend, in any way, on knowing values of μ; whatever μ may be, simple random sampling guarantees that the $\bar{x}$'s average out to it.

Example 11: As a very simple illustration, let us suppose that a finite population is made up of the 5 numbers 1, 2, 3, 4, 5, and suppose that simple random samples of size 2 are to be drawn. It can be easily shown that with simple random sampling, every sample of a given size has the same probability of being drawn. Assume it is desired to estimate the population mean μ (which we know to be 3) with the sample means. We will have:

Samples *	Values of $\bar{x}$	Associated Probability
(1, 2)	1.5	1/10
(1, 3)	2.0	1/10
(1, 4)	2.5	1/10
(1, 5)	3.0	1/10
(2, 3)	2.5	1/10
(2, 4)	3.0	1/10
(2, 5)	3.5	1/10
(3, 4)	3.5	1/10
(3, 5)	4.0	1/10
(4, 5)	4.5	1/10

*Order is not considered; for example the sample (1, 2) is the same as sample (2, 1).

We may consider $\bar{x}$ as a discrete random variable with the distribution given by:

Values of $\bar{x}$	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Prob. of value	1/10	1/10	2/10	2/10	2/10	1/10	1/10

$$\begin{aligned} \text{Thus, } E(\bar{x}) &= (1/10)[1.5 + 2.0 + 2(2.5) + 2(3.0) + 2(3.5) + 4.0 + 4.5] \\ &= (1/10)(30) = 3\ , \end{aligned}$$

so that $\bar{x}$ is unbiased estimate of μ.

It may be that for some population parameter, several different types of unbiased estimates are possible. In such a case, the one with the greatest precision would probably be the most desirable. We have seen that random sampling allows us to determine the expected value of estimates computed from samples of fixed size. It also allows us to determine their variances over all possible samples of a given size. The variance of estimates is used to measure their precision; larger variance means less precision, and vice versa. Thus, if we have several ways of unbiasedly estimating some unknown population quantity, we may choose those with smaller variances as being more desirable.

Example 12: For a population of size N with population variance σ^2, it can be shown that over all samples of size n, Var $(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N - n}{N - 1}\right)$. Thus, if there were some other unbiased estimate of μ besides $\bar{x}$ being considered, its precision relative to $\bar{x}$ could be determined by comparing its variance to $\frac{\sigma^2}{n}\left(\frac{N - n}{N - 1}\right)$. For the population of the previous example,

$$\begin{aligned} \sigma^2 &= \frac{1}{5}\sum_{1}^{5}(x_i - 3)^2 = \frac{1}{5}\ (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 \\ &= 2\ , \end{aligned}$$

so by formula,

$$\mathrm{Var}(\bar{x}) = \frac{2}{2}\left(\frac{5-2}{4}\right) = 0.75 \ .$$

By computation over all samples,

$$\mathrm{Var}(\bar{x}) = E(\bar{x}^2) - E^2(\bar{x}) \ ,$$

where

$$E(\bar{x}^2) = \frac{1}{10}\left[2.25 + 4.00 + (2)(6.25) + (2)(9.00) + 2(12.25) + 16.00 + 20.25\right] = 9.75$$

so that

$$\mathrm{Var}(\bar{x}) = 9.75 - 9.00 = 0.75 \ ,$$

in agreement with the formula.

For finite populations, it is customary to use $S^2 = \sum_1^N (y_i - \mu)^2/(N-1)$ rather than the variance σ^2 to measure the population variability, because both S^2 and σ^2 are equally valid measures of variability, and the use of S^2 permits some simplification. Since $S^2 = \frac{N}{N-1}\sigma^2$, we can write, $\mathrm{Var}(\bar{x}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$.

Although $\mathrm{Var}(\bar{x})$ can be computed and compared with variances of other estimates, the expression as it stands is of limited value, due to the fact that in most cases S^2 is an unknown quantity. However, S^2 may be also estimated from the sample, and

$$s^2 = \sum_1^n (x_i - \bar{x})^2/n - 1)$$

is an unbiased estimate of S^2 , that is, $E(s^2) = S^2$, where the expectation is taken over all samples of size n . Hence, an unbiased estimate of $\mathrm{Var}(\bar{x})$ is $\frac{s^2}{n}\left(1 - \frac{n}{N}\right)$.

The discussion to this point illustrates the type of activity associated with estimation in finite populations. It is important to note that there has been nothing, as yet, which indicates any connection between the mathematical distributions discussed earlier and inference in finite populations. In fact, the number of mathematical distributions which deal with finite discrete populations are few; indeed, the most common is the binomial distribution. Thus, the situation with finite populations is briefly this: unless the population definitely arises as the result of some process which specifically produces outcomes which follow a finite discrete mathematical distribution, or the population can be considered to be a subset of some infinite population, then statistical estimation is limited to finding estimates of unknown population quantities, such as the mean, variance, and median, and assessing the bias and precision of those estimates. Furthermore, these activities are possible only when random sampling is used and are consequences of the probability structure induced by random sampling.

Statistical testing, which is also an activity associated with inference, is similarly limited in finite populations. We shall illustrate one type of test by means of an example.

Example 13: A manufacturer of a medicine claims that at least 7 out of 10 doctors recommend his product for a particular ailment. To test this claim, suppose that from a random sampling of 100 M.D.'s in the United States, it was determined that 52 of them did recommend his product, Brand A, and 48 recommended something other than Brand A. On the basis of this sample, should the claim be considered extravagant? Since 100 doctors represents a very small proportion of the M.D. population, we may suppose that the sampling process did not noticeably change the relative proportions for or against Brand A. If the claim is valid, then at least 70% of the M.D.'s do recommend Brand A, and the process of sampling could then be considered as a binomial experiment with $n = 100$ and $p = 0.7$; i.e., 100 repetitions of a Bernoulli trial where P(Success) = P(Recommendation of Brand A) = 0.7. In such a case, we may ask what the probability is of getting 52 or less successes out of 100 trials. This is $\sum_{i=0}^{52} C_i^{100}(0.7)^i(0.3)^{100-i}$, approximately 0.0001. We reason as follows: either the manufacturer's claim is valid or it isn't. If it is valid, then the probability of getting a random sample of 100 M.D.'s with 52 or less positive responses is 0.0001. In other words, if, in fact, at least 7 out of 10 doctors do recommend Brand A, then of all the samples of size 100 which could be drawn, 99.99% of them would have yielded 53 or more Brand A recommendations, and what is more, 99.99% of all the samples would have yielded 64 or more Brand A recommendations. Thus, a claim of 70% is strongly contradicted by "experimental" evidence, and we would conclude that the proportion of M.D.'s recommending Brand A is almost certainly less than 70% and is probably somewhere between 40% and 60% .

The essential features of statistical testing are illustrated by this example. One has a tentative hypothesis about some aspect of the population, and he obtains observations from that population which tend either to support or to contradict the hypothesis. The degree of agreement or disagreement can be quantified whenever it is possible to construct from the observations some quantity which should have a certain behavior (distribution) if the tentative hypothesis does actually hold. One then asks whether or not the value of the quantity obtained from the observations could have plausibly occurred under the conditions imposed by the tentative hypothesis. If it is determined that the probability of getting such a value is small, then we say that the tentative hypothesis is contradicted, and we tend to regard it with suspicion, at the very least, and in many cases, to discard it altogether. The probability mentioned above is often referred to as the significance level , and statistical tests are sometimes called tests of significance or significance tests. They are also often called tests of hypothesis.

The area of statistical testing and decision-making has been a battleground for statistical theorists for many years, and a great amount has been written on these subjects. We would not presume to condense the matter into a few words, but we can say that much of the controversy has been concerned with whether the purpose of statistical testing should be regarded as one of decision-making or one of simply learning from the data. Also, a large portion of the material written about statistical testing has been concerned with finding optimal testing and/or decision procedures under various criteria of optimality.

The theoretical disagreements about statistical testing should not be permitted to obscure the fact that it has been found to be a highly useful statistical tool by researchers everywhere. Statistical testing helps one in forming opinions about populations and, moreover, assists one in objectively substantiating and defending those opinions. It is sometimes used as a device for decision-making and sometimes as a process of learning from the data, and it is a mistake to think that both functions are not served.

One should note that the construction of the test shown in Example 13 was made possible by two things: (1) random sampling, and (2) a large finite population. Actually, in most situations encountered, a population is large enough to be effectively considered an infinite one. In the next section, entitiled Statistical Inference in Infinite Populations, we shall discuss methods of estimation and testing which are appropriate whenever a population can be considered to be infinite and, for all practical purposes, continuous.

Although estimation and testing are generally thought of as two distinct and separate functions, there is a very useful technique of statistical inference which is sort of a hybrid between estimation and testing called consonance interval or confidence interval construction. Suppose we want to form an opinion about some parameter θ of a number population. A random sample of observations is drawn from the population, and from it an interval is calculated, within which the value of θ is believed to be. One way of interpreting such an interval is that it is a "list" of hypothetical values of the parameter θ which are consistent with the sample drawn. When this interpretation is used, the intervals are sometimes called consonance intervals.

Example 14: Assume that a population is large enough so that the drawing of one member at random does not essentially affect the probability that any other will be drawn. Actually, the use of the following requires the assumption of some nonspecific continuous underlying population distribution, but it can often be applied, in an approximate sense, to a large finite population. Suppose we want to find out something about the population median, and that a random sample of size 10 was drawn. Let the ordered sample be represented as $(x_1, x_2, \ldots, x_{10})$, and consider the question, "What is the probability that a sample would be drawn such that the population median lies between the smallest and largest observations of the sample?" Now, since the median is such that half of the population is on either side of it, any observation has a 50 - 50 chance of being greater or less than the median. Thus, the median would fail to be between the smallest and largest observations if all sample values were, respectively, all larger than the median, or all smaller than the median. The probability of this happening is $1/2^{10} + 1/2^{10}$ or $1/2^9$, about 0.002 . Thus, we would say that an estimated or a hypothesized value of the median less then x_1 or greater than x_{10} is highly inconsistent or disconsonant with the sample, and we could call the interval $(x_1 . x_{10})$ a 99.8% consonance interval for the median. Similar intervals with their consonance coefficients are given below.

(x_2, x_9)	97.9%
(x_3, x_8)	89.1%
(x_4, x_7)	65.6%
(x_5, x_6)	24.6%

Intervals given the consonance interpretation are probably best understood by considering the "degree of disconsonance" of values outside the interval.

Another way of looking at intervals constructed as above is regarding them as confidence intervals. Of all the (ranked) samples of size 10 that could be drawn, 99.8% of them will be such that the population median lies between x_1 and x_{10} . Hence, if one draws a sample of size 10 at random, and flatly claims that the population median is between x_1 and x_{10}, he has a 99.8% chance of being correct. Thus, we may call (x_1, x_{10}) a confidence interval on the population median with confidence coefficient 99.8% or simply a 99.8% confidence interval. Similar statements apply to the other intervals, the confidence coefficients being exactly identical to the consonance coefficients. It may seem tempting to discard a consonance statement or a confidence statement in favor of a simple probability statement such as, "The probability that the median lies between x_1 and x_{10} is 0.998," but we cannot do that, because the position of the median, although unknown, is fixed and is not a subject of probability. Furthermore, since the median is fixed, any given interval either will or will not include it; again, there is no probability involved. What the probability refers to is the method of making confidence statements, and is generally given a frequency interpretation such as: if one continued to draw random samples of size 10 and stated every time that the population median lay between the smallest and largest observation of the sample, then 99.8% of the time he would be making a correct statement.

Confidence intervals are most commonly applied to situations illustrated by the preceding example, to make statements about population quantities such as the mean, median, and variance. Two additional uses of confidence statements have been given special names. Sometimes it is useful to be able to make confidence statements about limits which are to include a given proportion of all population values, and such limits are called tolerance limits or tolerance intervals. Although the use of tolerance intervals also requires the assumption of some nonspecific continuous underlying population distribution, they too are often applied in an approximate sense to a large finite population. Another slightly different application of confidence statements is using them in connection with future observations. For example, for any given probability level p and any given proportion q, it is possible to construct limits which will include the proportion q of future observations with probability p. Confidence intervals used in this way are called prediction intervals. Tolerance intervals and prediction intervals will be discussed further in the section, "Statistical Inference in Infinite Populations."

We have introduced in this section three activities which can be considered the basic elements of statistical inference: estimation, testing, and construction of statistical intervals. These subjects will be discussed further in the next two sections.

5.3 Statistical Inference in Infinite Populations

The large majority of cases where one is attempting to learn something about a population by studying samples taken from it are such that the population of inference can be considered to be effectively infinite. In the case of finite populations, it was pointed out that no structure or distribution was assumed for the population; the only probability structure available for purposes of statistical inference was the result of random sampling. We also found that estimation, testing, and statistical interval construction were thereby somewhat limited, although the things which can be done represent a giant step above a policy of pure subjectivity. At this point, we will now suppose that a number population is such that its structure or behavior can be represented, or at least approximated with a probability distribution, and we shall use the terms population and distribution interchangeably. By and large, we will have in mind continuous distributions, although Poisson, geometric, negative binomial, and other infinite discrete distributions can describe certain infinite populations. Continuous populations are generally considered to arise as a result of numerical measurements. Classification or categorical populations are not of this type, and the methods of this section are not appropriate for studying them.

Whenever a population can be considered to follow some continuous distribution, it is not always necessary to know the specific form of the distribution. There are methods, called nonparametric statistical methods, for which it is sufficient to assume only that the population follows some continuous distribution without specifying the actual form. The other methods, which have been by far the most studied, most applied, and most discussed, are the so-called parametric methods, whereby the form of the distribution is specified except for one or more unknown parameters. The parametric approach, surprisingly enough, still allows a great deal of latitude. The normal family of distributions, for example, with μ and σ^2 allowed to vary, encompasses a wide variety of shapes and locations. However, the primary reason for the prevalence of parametric methods in present-day statistical methodology is the fact that a great body of powerful mathematical theory may be brought to bear with regard to determining optimum procedures: "best" estimates, "most powerful" tests, etc. It is possible to answer questions like, "Of all the unbiased estimates of some population parameter, does one exist having maximum precision (minimum variance), and if so, how may it be found?"

Random sampling from infinite populations translates into, "drawing observations independently." From a theoretical standpoint, it is convenient to think of a sample of n observations drawn at random as being a set of n independent random variables, each having the same distribution as the population from which it was drawn. Quantities calculated from samples, such as a sample mean or sample variance, will then be functions of random variables and so will be themselves random variables having distributions. Such quantities are called statistics.

Traditionally, the term statistic has been applied to any descriptive quantity calculated from a set of numbers; however, when we speak here of a statistic, we shall mean a random variable which is a function of sample observations and which does not depend upon any unknown parameters.

Example 15: Let a sample of size n be drawn at random from a normal distribution with mean μ and variance σ^2; denote the sample by $(x_1, x_2, \ldots, x_n)$. Then, the sample mean $\bar{x} = \sum_1^n x_i/n$ and the sample variance $s^2 = \frac{1}{n-1} \sum_1^n (x_i - x)^2$ are statistics, and it can be shown that $\bar{x}$ has a normal distribution with mean μ and variance σ^2/n, and that $(n-1)s^2/\sigma^2$ has a Chi-Square distribution with $n-1$ degrees of freedom. The purpose of considering these statistics in the first place is that $\bar{x}$ and s^2 are unbiased estimates of μ and σ^2, respectively; i.e., $E(\bar{x}) = \mu$ and $E(s^2) = \sigma^2$. A somewhat startling fact is that $\bar{x}$ and s^2 are independent random variables whenever the parent distribution is normal. One would probably not have suspected that this could be possible, as the formula for s^2 seems to functionally involve $\bar{x}$. A useful algebraic identity is

$$\sum_1^n (x_i - \bar{x})^2 = \sum_1^n x_i^2 - n\bar{x}^2 = \sum_1^n x_i^2 - \frac{\left(\sum_1^n x_i\right)^2}{n},$$

and one of the two expressions on the right is generally used to actually compute s^2.

Now, suppose that the value of μ is known, say $\mu = \mu_0$. Since $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ would then be a standard normal variable and $(n-1)s^2/\sigma^2$ an independent Chi-Square variable, we may form

$$t = [(\bar{x} - \mu_0)/(\sigma/\sqrt{n})]/\sqrt{(n-1)s^2/(n-1)\sigma^2} = \sqrt{n}(\bar{x} - \mu_0)/s,$$

which has the Student's t-distribution with $n-1$ degrees of freedom.

It is useful to think of the t-distribution in relation to a normal distribution. If X has a normal distribution with mean zero and variance σ^2, then X/a has the standard normal distribution. Suppose that we do not know σ^2, but can unbiasedly estimate it from n observations with the statistic s^2. Then X/s has the t-distribution with n-1 degrees of freedom, the distribution depending only upon the number of observations available for estimating σ^2. The shape of the distribution of X/s is similar to that of X/σ, except the distribution of X/s is more spread out, owing to the variability of the statistic s. As n increases, the t-distribution becomes more like the standard normal until, for 100 degrees of freedom, a t-distribution is, for all practical purposes, the same as a standard normal distribution.

There are some facts about continuous distributions hinted at in the preceding example, across almost all continuous distributions. Practically every continuous distribution of interest has a mean μ and a variance σ^2, expressible in terms of distribution parameters. In random samples of size n from any of these distributions:

1. The sample mean $\bar{x}$ is an unbiased estimate of μ, and $\mathrm{Var}(\bar{x}) = \sigma^2/n$.
2. The sample variance s^2 is an unbiased estimate of σ^2.
3. An unbiased estimate of $\mathrm{Var}(\bar{x})$ is $s^2_{\bar{x}} = s^2/n$.

Thus, the precision of the sample mean as an estimate of the population mean depends on the sample size and the inherent variability of the population. Since $\mathrm{s.d.}(\bar{x}) = \sigma_{\bar{x}} = \sigma/\sqrt{n}$, it is evident that for a given population, the standard deviation of a sample mean is inversely proportional to the square root of the sample size. Precision is often expressed in terms of standard deviation instead of variance, and when it is, the statement is often made that to increase the precision of a sample mean by a factor of k, the sample size must be increased by a factor of k^2. For example, to reduce the standard deviation of a sample mean by a factor of 1/2 (double the precision), the sample size must be increased by a factor of 4.

5.4 Statistical Inference in Normally Distributed Populations

As previously mentioned, many statistical techniques are based on the assumption that the population of inference follows a normal distribution. To reiterate briefly, mathematical tractability, usefulness for approximations, and consequences of the Central Limit Theorem were suggested as justifications for using the normal distribution. Additionally, we should note that in several instances in the sciences where probabilistic models have been used to describe physical phenomena, theoretical considerations have led directly to a normal distribution. In many more cases, it can be shown that a normal distribution approximately fits the situation and, often, an approximate population distribution is quite satisfactory. The general rule of operation whenever a continuous distribution is thought to describe some population is this: if there is good reason to assume that some specific continuous distribution applies (normal, exponential, Chi-Square, etc.), then the appropriate theory for the particular distribution is employed. If the form of the distribution cannot be determined then, unless there is definite evidence to the contrary, a normal distribution is often used in an approximate sense. If the form of the distribution is not known, and it is definitely felt that the assumption of a normal distribution is not warranted, then one usually reverts to one of the several nonparametric methods. Only one or two nonparametric methods have been or will be discussed, but they are generally straightforward and uncomplicated to apply. One who is interested in nonparametric methods may find them discussed in references.

In the following discussions, we shall assume that a normal population distribution applies, but that the parameters μ and σ^2 are unknown and that inferences are to be made concerning them. We shall restrict ourselves to the basic elements of statistical inference: estimation, testing, and construction of statistical intervals.

5.4.1 A Single Normal Population

Suppose that a random sample of size n, $(x_1, x_2, \ldots, x_n)$ is drawn from a normal population with unknown mean μ and unknown variance σ^2. Then we know that the following facts are true:

1. The sample mean $\bar{x} = (\Sigma x_i)/n$ is an unbiased estimate of μ, and over repeated sampling, $\bar{x}$ follows a normal distribution with mean μ and variance σ^2/n. The variance of $\bar{x}$ is smaller than any other unbiased estimate of μ, so that $\bar{x}$ is the unbiased estimate of μ with maximum precision. We characterize this situation by saying that $\bar{x}$ is a minimum-variance unbiased estimate of μ. If a minimum-variance unbiased estimate is unique then we say that it is the best unbiased estimate. Thus, $\bar{x}$ is the best unbiased statistic for estimating μ.
2. The sample variance

$$s^2 = \frac{1}{n-1}\left(\Sigma x_i^{\,2} - \frac{(\Sigma x_i)^2}{n}\right)$$

is the best unbiased estimate of σ^2, and over repeated sampling, $(n-1)s^2/\sigma^2$ has a Chi-Square distribution with n-1 degrees of freedom.

3. The statistic s^2/n is the best unbiased estimate of $\sigma_{\bar{x}}^2$.
4. For any μ_0, if the true value of μ is μ_0, then $\sqrt{n}(\bar{x}-\mu_0)/s$ is distributed as Student's t with n-1 degrees of freedom. That is, over repeated sampling with sample size n, with a new $\bar{x}$ and s^2 calculated for each new sample, repeated calculations of $\sqrt{n}(\bar{x} - \mu_0)/s$ would yield numbers occurring according to the probability law for a Student's t-distribution with n-1 degrees of freedom.

Suppose we suspect that the value of μ is μ_0. Then we say that the tentative null hypothesis is that $\mu = \mu_0$, and we may test the tenability of μ_0 by taking a random sample $(x_1, x_2, \ldots, x_n)$, and from it calculating $t_{cal} = \sqrt{n}(\bar{x} - \mu_0)/s$. Now, if μ_0 were the exact value of μ , then t_{cal} should be a reasonably occurring value of a random variable having a Student's t - distribution with n-1 degrees of freedom. Suppose, however, that we are mistaken, that μ is really $\mu_1 < \mu_0$. Then, in this case, $t_1 = \sqrt{n}(\bar{x} - \mu_1)/s$ would have been the proper value to calculate. We see that what we actually calculated, t_{cal}, and what we should have calculated, t_1, are related by

$$\begin{aligned} t_{cal} = \sqrt{n}(\bar{x} - \mu_0)/s &= (\sqrt{n}(\bar{x} - \sqrt{n}\mu_1 + \sqrt{n}\mu_1 - \sqrt{n}\mu_0)/s \\ &= \sqrt{n}(\bar{x} - \mu_1)/s - \sqrt{n}(\mu_0 - \mu_1)/s \\ &= t_1 - \sqrt{n}(\mu_0 - \mu_1)/s \quad . \end{aligned}$$

Thus, if μ is really μ_1, some value less than μ_0, t_{cal} is a number less than the Student's t-value t_1 by an amount $n(\mu_0 - \mu_1)/s$. In such a case, we would get a value for t_{cal} which would tend to be too low a value for a reasonable t-variable to assume, and the greater the difference $\mu_0 - \mu_1$, the lower t_{cal} would be in relation to a reasonable t-value. If, on the other hand, μ is some value $\mu_2 > \mu_0$, then t_{cal} is larger than a Student's t-variable by an amount $\sqrt{n}(\mu_2 - \mu_0)/s$.

Therefore, to test the tenability of the value of μ_0 as the population mean, we calculate $t_{cal} = \sqrt{n}(x - \mu_0)/s$, and determine if t_{cal} is a reasonable value for a Student's t-variable with n-1 degrees of freedom to assume. If t_{cal} is judged to be too large or too small, then we say that there is strong evidence against the hypothesis that $\mu = \mu_0$.

Throughout this discussion, we have used the word "reasonable" without explanation. What is meant by a "reasonable" value of a Student's t-variable? To answer this question, we should examine a table of the Student's t-distribution. Such a table can be found in a CRC Handbook of Chemistry and Physics. For the various degrees of freedom, one will find critical values tabulated under columns for p-values ranging from 0.9 to 0.01. The p-values refer to two-tailed critical values, and the meaning of the terms is as follows: consider the tabulated values for, say, 10 degrees of freedom. For the two-tailed p-value 0.5, the critical value of t is given as 0.7. This means that $Pr(-0.7 < t < 0.7) = 0.5$, or 50% of the t-values fall between -0.7 and 0.7. Reading across, we similarly find that 60% are between -0.879 and 0.879, since $P(-.879 < t < .879) = 0.6$, ..., 95% are between -2.228 and 2.228, and so on. Often, tail areas called α-levels are sometimes needed. To find the two-tailed critical value for a given α-level, one finds the corresponding two-tailed critical value for $p = 1 - \alpha$. Also, it is sometimes necessary to know the probability of getting a value larger than a certain value, say t_0, or the area in the right tail of the t-distribution above t_0 . Since the t-distribution is symmetric, the one-tail α-level is $(1 - p)/2$, where $p = \Pr(-t_0 < t < t_0)$ as given by the table.

Now, suppose that from a sample of size 10, $t_{cal} = \sqrt{n}(\bar{x} - \mu_0)/s$ was calculated to be 2.8; is this a reasonable t-value, or is it too large? From the t-table for 9 degrees of freedom, we determine that the probability of getting a t-value larger than 2.8 and that, if μ really has the value μ_0, then on the average, only one sample in a hundred will give a t_{cal} of 2.8 or larger. What if the sample had given a t_{cal} of 0.8 instead of 2.8? In this case, the probability of a t-value larger than 0.8 is about 0.22 . Should a t-value of 0.8 be considered unusual? It should now be evident that "reasonableness" is largely a matter of personal preference. The probabilities 0.01 and 0.22 are called significance levels. The smaller the significance level, the stronger the evidence against the null hypothesis, but the strength of evidence needed to discredit the tentative null hypothesis varies from person to person and from situation to situation. As a general rule, people tend to consider a significance level of 0.05 or less as strong evidence against the null hypothesis.

It is also possible to construct confidence intervals for the parameters μ and σ^2 .

(a) Confidence Interval for μ

Since $\sqrt{n}(\bar{x} - \mu)/s \sim t(n - 1)$, for any level of confidence desired, say 90%, we may find from a t-table, a value t_0 such that $\Pr(-t_0 < t(n-1) < t_0) = 0.9$, and we may assert (before any one sample is drawn) that $\Pr(-t_0 < \sqrt{n}(\bar{x} - \mu)/s < t_0) = 0.9$. Changing quantities around, we may make the above interval into a fixed-length interval with random placement for which the probability statement still holds, that is,

$$\Pr(-t_0 < \sqrt{n}(\bar{x} - \mu)/s < t_0) = 0.9$$

so $$\Pr(-t_0 s/\sqrt{n} < x - \mu < t_0 s/\sqrt{n}) = 0.9,$$

and $$\Pr(\bar{x} - t_0 s/\sqrt{n} < \mu < \bar{x} + t_0 s/\sqrt{n}) = 0.9.$$

Thus, we think of the limits $x \pm t_0 s/n$ as a random interval, and over repeated sampling, intervals will be generated, 90% of which will contain μ. After taking a sample, we have no way of knowing whether the resulting interval does or does not contain μ. However, we do know that only 10% of all the intervals which could be constructed in this way would fail to include μ and, accordingly, we call $\bar{x} \pm t_0 s/\sqrt{n}$ a 90% confidence interval for μ.

(b) Confidence Interval for σ^2

Suppose a 95% confidence interval is desired. Since $(n - 1)s^2/\sigma^2 \sim \chi^2(n - 1)$, we may find from tables of Chi-Square values a_1 and a_2 such that $\Pr((n - 1)s^2/\sigma^2 < a_1) = 0.025$, or $\Pr(a_1 < (n - 1)s^2/\sigma^2 < a_2) = 0.95$ (a_1 and a_2 must be found separately because of the asymmetry of the Chi-Square distribution). From the probability statement above, we derive $[(n - 1)s^2/a_2, (n - 1)s^2/a_1]$ as a 95% confidence interval on σ^2.

Example 16: For a pilot-plant experiment, 500 gallons of reactant must be mixed prior to a day's run. Although some leeway is permissible, for best results, the pH of the solution should not exceed 6.3 by much and should not be much below 6.1. Experience has shown that fluctuations in measurements of this type tend to be normally distributed. Let us assume that this is the case here. Over an hour's time, 8 samples were taken yielding pH values 6.23, 6.31, 6.29, 6.19, 6.23, 6.17, 6.15, 6.21 . On the basis of this sampling, may the experiment be started? We calculate

$$\bar{x} = 6.22, \; s^2 = 0.0031, \; s/\sqrt{n} = 0.0196$$

$$\frac{\sqrt{n}(\bar{x} - 6.3)}{0.0196} = \frac{-0.08}{0.0196} = -4.08 \text{ and } \frac{\sqrt{n}(\bar{x} - 6.1)}{s} = \frac{0.12}{0.0196} = 6.12 \, .$$

If the actual pH is 6.3 or higher, the probability of getting a sample yielding a t-value as low as -4.08 is small, less than 0.005. Thus, we would say that a pH of 6.3 or higher is strongly contradicted by the sample. Similarly, if the actual pH is 6.1 or lower, the probability of getting a sample with a t-value as high as 6.12 is less than 0.0005, so a pH lower than 6.1 is also contradicted by the data.

A 99% confidence interval on the pH of the solution is 6.22 ± (3.499)(0.0196), or 6.15 to 6.29. Using the consonance interpretation for the confidence interval, we could have combined the testing into the single statement that a true solution pH lower than 6.15 or higher than 0.29 is highly inconsistent with the sample evidence.

A confidence interval on a parameter does not usually give much information about population members, and generally should not be used for this purpose. However, statistics can be calculated from a sample which may be used in constructing confidence intervals for population members.

Let us first consider tolerance intervals, which are confidence statements about population proportions. Let us also assume that the parent population is normally distributed with unknown mean and variance. From a random sample of size n, calculate the sample mean $\bar{x}$ and the sample standard deviation s . Suppose we desire an interval for which we can be $100\gamma\%$ confident that it includes at least 100P% of the population. Special tables are available giving tolerance factors $K(n,\gamma,P)$ for various combinations of sample sizes n, confidence levels γ, and coverages P. From such a table, find the appropriate factor, say K; then $\bar{x} \pm Ks$ defines a $100\gamma\%$ tolerance interval for 100P% of the population.

Example 17: For the data of Example 16, let us calculate a 95% tolerance interval for 90% of all pH readings obtained by the same method. Recall that $\bar{x} = 6.22$ and $s = 0.0554$. From a CRC Handbook of Tables for Probability and Statistics (6), we find in the Table of Tolerance Factors for Normal Distributions, $K(8, 0.95, 0.90) = 3.264$. Thus, $6.22 \pm (3.624)(0.0554)$ or 6.22 ± 0.1808 defines the desired tolerance interval, and we say that with 95% confidence, 90% of all pH readings which could be obtained in the same way as the sample of 8 are between 6.04 and 6.40.

It is possible to construct distribution-free or nonparametric tolerance limits whenever the assumption of a specific underlying continuous distribution for the parent population is not justified; however, this case will not be discussed. For a discussion of distribution-free tolerance intervals, a standard test on nonparametric methods such as Conover (9) should be consulted.

5.4.2 Two Non-Independent Normal Populations

A very common and effective method of experimentation is the so-called comparative experiment. It is used whenever two sets of observations are thought to be correlated with each other. For example, an effective way to test the effect of a new feed ration for animals is to select pairs of individuals from several litters and to give one member of each pair the new ration. In this type of experiment, it is the difference in response that is of interest. The methods of analysis for the paired or comparative experiment are applicable to any case involving two samples, whether they are correlated or not, but the use of comparative experimentation is most effective whenever pairs of "similar" individuals or objects are used.

The discussion will be for the case when the respective populations from which the pairs are drawn are assumed to be normally distributed. If normality assumptions do not appear to be warranted, a comparative experiment can still be analyzed with one of the non-parametric techniques such as the sign test, the Wilcoxon sign-rank test, or the Walsh test (see Siegel (31)).

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of observations drawn at random from normal populations with the x's having a normal (μ_1, σ_1^2) distribution and the y's having a normal (μ_2, σ_2^2) distribution. Also assume that each x_i and y_i is correlated, with correlation equal to ρ. Note that because of random sampling, there will be no correlation between x_i and x_j, y_i and y_j, or x_i and y_j for $i \neq j$. Form a "new" sample of differences $(d_1, d_2, \ldots, d_n)$ by taking $d_i = x_i - y_i$. Then the following facts hold:

1. If the x's and y's are normally distributed, then the d's are also normally distributed with mean $\mu_d = \mu_1 - \mu_2$ and variance $\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.
2. Over repeated sampling, the sample mean $\bar{d}$ has a normal distribution with mean $\mu_{\bar{d}} = \mu_1 - \mu_2$ and variance $\sigma_{\bar{d}}^2 = \sigma_d^2/n$. The best unbiased estimate of $\mu_1 - \mu_2$ is $\bar{d}$.
3. Over repeated sampling, the sample variance

   $$s_d^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\Sigma d_i)^2}{n}\right]$$ is the best unbiased estimate

   of σ_d^2 , and $(n-1)s_d^2/\sigma_d^2$ has a Chi-Square distribution with n-1 degrees of freedom.
4. An estimate of the standard deviation of $\bar{d}$, $\sigma_{\bar{d}}$ is $s_{\bar{d}} = \sqrt{s_d^2/n}$.

Suppose an experiment yielded a positive $\bar{d}$, and suppose that a real positive difference $(\mu_1 - \mu_2)$ of magnitude d would be of practical significance. We may ask if the experiment gives strong evidence that the difference is, in fact, greater than d. To test the hypothesis, form

$$t_{cal} = \frac{\bar{d} - d}{s_{\bar{d}}}$$ based on n-1 degrees of freedom .

We then compare t_{cal} with values in a t-table for n-1 degrees of freedom. The numbers given are, of course, values in the distribution of a Student's t random variable with n-1 degrees of freedom; call the random variable t(n-1). Then, if $\Pr(t(n-1) > t_{cal})$, as determined from the table, is small, we would say that there is strong evidence that there is a real difference of at least d .

To get a 100p% confidence interval on the true difference, from the t-table for n-1 degrees of freedom, find t_0 such that $Pr(-t_0 < t(n-1) < t_0) = p$. Then $\bar{d} \pm t_0 s_{\bar{d}}$ defines the 100p% confidence interval.

Example 18: In a coal-fired boiler, Unit 10 at the Shawnee Steam Plant, Paducah, Kentucky, there are essentially two separate furnaces, the East side and the West side. During limestone injection testing on the unit, although input and control variables for the two sides were thought to be approximately the same, with respect to some variables the two sides seemed to be consistently different. Without trying to explain the differences, it was decided to try to determine, on the basis of the variation observed in the variables, whether or not there was any evidence that real differences existed. The complete data will not be given. The number of paired observations available for testing each variable varied, but where readings were taken for both East and West sides, a difference d_i = (East reading) - (West reading) was formed. Thus, for each variable, a sample of d_i's was formed, a $\bar{d}$ calculated, and $s_{\bar{d}}$ estimated. With the assumption that the variation in the variables is governed by normality, t-values such as $t_{cal} = \bar{d}/s_{\bar{d}}$ may be calculated and tested. The summary of results is given below:

Variable	$\bar{d}$	$s_{\bar{d}}$	t_{cal}	df	Conclusions*
Initial SO_2 Concentration	32.3	14.7	2.2	100	East >> West
Plane AA Temperature	-9.9	9.2	-1.1	63	East ≈ West
% Limestone Utilization	-.97	.19	-5.0	97	East << West
% Excess Air	.88	.52	1.7	95	East ≈ West
% Hydration	-1.8	.59	-3.4	45	East << West
% SO_2 Removal	-2.0	.5	-4.0	97	East << West

*>>(<<) means "is significantly higher (lower) than"
≈ means "is not significantly different from"

5.4.3 Two Independent Normal Populations

Often when two samples are taken, there is no reason to consider them related or dependent. Let $(x_1, x_2, \ldots, x_k)$ and $(y_1, y_2, \ldots, y_m)$ be respective random samples from normal populations with $x_i \;\; N(\mu_1 . \sigma_1^2)$ and $y_i \;\; N(\mu_2, \sigma_2^2)$. Then we have:

1. $\bar{x}$ is the best unbiased estimate of μ_1; $\bar{y}$ is the best unbiased estimate of μ_2; and over repeated sampling, $\bar{x} \sim N(\mu_1, \sigma_1^2/k)$ and $\bar{y} \sim N(\mu_2, \sigma_2^2/m)$.
2. $\bar{x} - \bar{y}$ is the best unbiased estimate of $\mu_1 - \mu_2$, and over repeated sampling, $(\bar{x} - \bar{y}) \sim N(\mu_1 - \mu_2, \sigma_1^2/k + \sigma_2^2/m)$.
3. $s_1^2 = \frac{1}{k-1}[\Sigma x_i^2 - (\Sigma x_i)^2/k]$ and $s_2^2 = \frac{1}{m-1}[\Sigma y_i^2 - (\Sigma y_i)^2/m]$ are, respectively, the best unbiased estimates of $\mu^2{}_1$ σ_2^2 .
4. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, then the best unbiased estimate of σ^2 is a weighted combination of the estimates s_1^2 and s_2^2. The combined estimate is called a pooled estimate and is given by

$$s^2 = \frac{(k-1)s_1^2 + (m-1)s_2^2}{k + m - 2} = \frac{[\Sigma x_i^2 - (\Sigma x_i)^2/k] + [\Sigma y_i^2 - (\Sigma y_i)^2/m]}{k + m - 2} .$$

5. An estimate of the standard deviation of $\bar{x} - \bar{y}$,

$$\sigma_{\bar{x} - \bar{y}} \text{ is } \sqrt{\frac{s_1^2}{k} + \frac{s_2^2}{m}} \text{ if } \sigma_1^2 \neq \sigma_2^2 , \text{ and}$$

$$\sqrt{\frac{1}{k} + \frac{1}{m}} \cdot s \quad \text{if } \sigma_1^2 = \sigma_2^2 .$$

It is very important in the analysis of two independent samples to know whether or not σ_1^2 and σ_2^2 can be considered to be the same. This can be tested with an F-test for equality of variances as follows:

Suppose s_1^2 is somewhat greater than s_2^2 so that we may suspect that σ_1^2 is greater than σ_2^2. Now, $(k-1)s_1^2/\sigma_1^2$ and $(m-1)s_2^2/\sigma_2^2$ are distributed as Chi-Squares with k-1 and m-1 degrees of freedom, respectively. Recall that a ratio of Chi-Square random variables divided by their degrees of freedom has an F distribution; thus,

$$\frac{(s_1^2/\sigma_1^2)}{(s_2^2/\sigma_2^2)}$$ has an F distribution with k-1 and m-1 degrees of freedom.

If $\sigma_1^2 = \sigma_2^2$, then the variances cancel out of the expression. Therefore, if $\sigma_1^2 = \sigma_2^2$, then s_2^2/s_2^2 should be a reasonably occurring value of an F random variable with k-1 and m-1 degrees of freedom. We may compare the value of s_1^2/s_2^2 obtained with tabulated values for an F(k-1, m-1). Such a table can be found in any applied statistics text and in a CRC Handbook of Chemistry and Physics. If s_1^2/s_2^2 is "too large," then it is concluded that σ_1^2 is probably larger than σ_2^2; otherwise, we say that there is insufficient evidence to consider σ_1^2 and σ_2^2 to be different. Depending upon the results of the F-test, we can take one of two alternatives:

Case 1: Variances Assumed the Same

If one concludes that σ_1^2 and σ_2^2 are probably about the same, then s_1^2 and s_2^2 can be regarded as two independent estimates of the same thing, and they can be combined as in 4, above, for a pooled estimate s^2, which makes use of all the observations.

Suppose an experiment yielded a positive difference $\bar{x} - \bar{y}$, and suppose a real positive difference of magnitude d is practically significant. To test whether there is evidence that μ_1 is greater than μ_2 or d or more, form

$$t_{cal} = \frac{(\bar{x} - \bar{y}) - d}{\sqrt{\frac{1}{k} + \frac{1}{m}}\, s} \text{ based on } k + m - 2 \text{ degrees of freedom,}$$

and compare t_{cal} with tabulated values of $t(k + m - 2)$. If $Pr(t(k + m - 2) > t_{cal})$ is small, then we say that there is evidence of a real difference of at least d.

A 100p% confidence interval on $\mu_1 - \mu_2$ is given by $(\bar{x} - \bar{y}) \pm t_0\sqrt{\frac{1}{k} + \frac{1}{m}}\, s$, where t_0 is a value from the t-table for $k + m - 2$ degrees of freedom such that $Pr(-t_0 < t(k + m - 2) < t_0) = p$.

Case 2: Variances Assumed Different

Unfortunately, whenever $\sigma_1^2 \neq \sigma_2^2$, $\sqrt{\frac{s_1^2}{k} + \frac{s_2^2}{k}}$ does not form the denominator of a t random variable except in an approximate way. The problem of testing and constructing confidence intervals for the case $\sigma_1^2 \neq \sigma_2^2$ is one of some note and is known as the Behrens-Fisher problem. There have been several satisfactory approximate methods proposed; the following one is called the Aspin-Welch solution:

Let s_1^2/k be denoted by $s_{\bar{x}}^2$, let s_2^2/m be denoted by $s_{\bar{y}}^2$, and let $s_{\bar{x}}^2 + s_{\bar{y}}^2 = s_1^2/k + s_2^2/m$ be denoted by s_0^2. Then, if the true difference $\mu_1 - \mu_2$ is d,

$$t' = \frac{(\bar{x} - \bar{y}) - d}{s_0}$$

has an approximate t distribution with r degrees of freedom, where r is computed by

$$\frac{1}{r} = \frac{1}{k-1}\left(\frac{s_{\bar{x}}^2}{s_0^2}\right) + \frac{1}{m-1}\left(\frac{s_{\bar{y}}^2}{s_0^2}\right)^2 .$$

Testing and confidence interval calculations are similar to Case 1, except t' and s_0 are used.

Example 19: (Snedecor (32))

A quick but imprecise method of estimating the concentration of a chemical in a vat has been developed. Eight samples from the vat are analyzed by the quick method, and four samples are analyzed by the standard method which is precise but slow. It is desired to determine whether results by the quick method are biased. The data and calculations are:

Standard Analysis (x)	Quick Analysis (y)
25	23
24	18
25	22
26	28
	17
	25
	19
	16

$\bar{x} = 25$

$s_1^2 = 0.67$

$s_{\bar{x}}^2 = s_1^2/4 = 0.17$

$\bar{y} = 21$

$s_2^2 = 17.71$

$s_{\bar{y}}^2 = s_2^2/8 = 2.21$

$$s_0^2 = 2.38$$

$$\frac{1}{r} = \frac{1}{3}\left(\frac{0.17}{2.38}\right)^2 + \frac{1}{7}\left(\frac{2.21}{2.38}\right)^2 = .1249$$

$$r = 8$$

First, it appears that s_1^2 and s_2^2 are not estimating the same thing. $F_{cal} = 17.71/0.67 = 26.43$ with 7 and 3 degrees of freedom supports this, as the probability of an F-value as large as 26.43 occurring (if σ_2^2 is no larger than σ_1^2) is less than 0.025 and only slightly more than 0.01 . Thus we should assume that $\sigma_1^2 \neq \sigma_2^2$ and use t' for testing and confidence intervals. Let us evaluate the probability of getting a difference of $\bar{x} - \bar{y} = 4$ if the quick method is not biased. We use

$$t'_{cal} = 4/\sqrt{2.38} = 2.6 \text{ with } r = 8 \text{ degrees of freedom.}$$

The probability of getting a value as large as 2.6 (if $\mu_2 - \mu_1 = 0$) is less than 0.17 . Since this is small, there is strong evidence that the quick method of analysis consistently underestimates the concentration.

Let us take this opportunity to illustrate the use of a one-sided confidence interval. We shall find a number which we are 90% confident that $\bar{x} - \bar{y}$ exceeds; that is, a lower 90% confidence limit for the difference $\mu_1 - \mu_2$. From a t-table for 8 degrees of freedom, we must find a value t_0 such that $Pr(t(8) < t_o) = 0.9$. The value of t_0 is 1.397, so that $(\bar{x} - \bar{y}) - t_0 s_{\bar{x}-\bar{y}} = 4 - (1.397)(2.38) = 1.84$ is the desired limit. Hence, we say that we are 90% confident that the difference $\mu_1 - \mu_2$ is at least 1.84 units.

5.4.4 Several Independent Normal Populations

For more than two populations, we will consider only the simplest case. Let $(x_{11}, x_{12}, \ldots, x_{1n})$, $(x_{21}, x_{22}, \ldots, x_{2n})$, ..., $(x_{k1}, x_{k2}, \ldots, x_{kn})$ be k samples all of size n drawn from normal populations such that for any $i = 1, 2, \ldots, k$. $x_{ij} \sim N(\mu_i, \sigma^2)$ for all $j = 1, 2, \ldots, n$. That is, all k samples are from normal populations having the same variance σ^2 but possibly different means μ_i. Stated in general terms, the goal is to draw inferences of some sort about the means μ_i. Translated into specific terms, one or more of the following questions (as well as some others) might be asked:

1. Are there any differences in the μ_i, and if so, where are the differences?
2. Can some grouping pattern in the μ_i be determined?
3. What can be said about some specific differences, or combinations of differences?
4. How may confidence intervals be constructed for some or all of the μ_i and for some or all of the differences $\mu_i - \mu_j$?

The answers to these and other similar questions have been attempted by the development of multiple inference procedures. The problem of multiple inferences has been kicked around for years, and the theory is by no means complete on the subject. One of the basic difficulties seems to center around how to deal with error rates. The situation is most easily explained by using confidence intervals. Assume that one adopts a fixed confidence level, say 99%, for constructing all his confidence intervals; then we would say that his error rate is 1%, because on the average, 1% of all the intervals he could construct would fail to cover the parameter of interest, whenever he continued to take independent samples from the same population. Moreover, if he constructed N confidence intervals, he should expect (0.01)N of them to be wrong. However, what seems to bother people is the fact that with an error rate of α, the probability of getting k or more independent confidence intervals without an error is less than $1 - \alpha$ and is actually $(1 - \alpha)^{k-1}$. For example, with $\alpha = 0.05$, the probability of getting 10 or more independent confidence intervals without an error is only about 0.6 . Ordinarily, people don't get too upset about it and accept as a "fact of life" that regardless of how unlikely an event may be, if "fate" is tempted long enough, the event is eventually almost sure to occur. Perhaps, the difficulty is largely psychological, that there is a natural reluctance to "push one's luck too far" at any one time. In spite of the theoretical controversy surrounding multiple inference procedures, they are considered by most to be useful techniques and are, in fact, widely used.

We will not consider question 5 in any detail, but will try to show what may be done to answer 1, 2, and 3. Let us consider 3 first. Whenever more than two means are involved, the concept of difference can be generalized by the use of contrasts or comparisons. Consider a linear combination of population means $r_1\mu_1 + r_2\mu_2 + \ldots + r_k\mu_k$ such that $\Sigma r_i = 0$. Such a linear combination can be estimated unbiasedly by the same linear combination of sample means $r_1\bar{x}_1 + r_2\bar{x}_2 + \ldots + r_k\bar{x}_k$, where $\bar{x}_i$ is the sample mean of the ith sample $(x_{i1}, x_{i2}, \ldots, x_{in})$, and any linear combination $\Sigma r_i\bar{x}_i$ of sample means is called a contrast or comparison on the means if $\Sigma r_i = 0$ with not all $r_i = 0$. Let $C = \Sigma r_i\bar{x}_i$ be a contrast on sample means. Then:

1. Over repeated sampling, C has a normal distribution with mean $\Sigma r_i\mu_i$ and variance $(\Sigma r_i^2)\sigma_i^2/n_i$ for unequal sample sizes and unequal variances.
2. The best estimate of Var(c) is $s_c^2 = (\Sigma r_i^2)s_0^2/n$, where s_0^2 is the pooled estimate of σ^2 based on k(n-1) degrees of freedom; i.e.,

$$s_0^2 = \frac{(n-1)[s_1^2 + s_2^2 + \ldots + s_k^2]}{k(n-1)} = \sum_{i-1}^{k} s_i^2/k, \text{ for}$$

s_i^2 = sample variance of the ith sample =

$$\frac{1}{n-1}[\sum_j x_{ij}^2 - \sum_j x_{ij}{}^2/n] \ .$$

3. Let $C_0 = \Sigma r_i \mu_i$, then $(C - C_0)/s_C$ is distributed as a Student's t random variable with k(n-1) degrees of freedom.

Note that a contrast may involve less than all sample means, since some of the r_i are not prevented from being zero. Thus, a difference $\bar{x}_i - \bar{x}_j$ is a contrast, and the estimate of the standard deviation of $\bar{x}_i - \bar{x}_j$ can be seen to be $\sqrt{2/n}\ s_0$ based on k(n-1) degrees of freedom. Thus, for any general contrast C, $t = (C - C_0)/s_c$ is appropriate for testing and confidence intervals, and for contrasts which are simple differences, $t = [(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)]/\sqrt{2/n}\ s_0$ is the appropriate quantity. For example, to test whether there is evidence that $\mu_1 \neq \mu_2$, form $t_{cal} = (\bar{x}_1 - \bar{x}_2)/\sqrt{2/n}\ s_0$, and compare with tabulated t(k(n-1)).

One of the problems associated with testing only certain contrasts is that an experimenter is sometimes told that the contrasts to be tested must have been planned before the data was taken and that making unplanned tests is to be discouraged. There are theoretical justifications for this advice, but it is often puzzling to someone to be told that the statistical significance of his results depends upon whether he planned to find them before the experiment or not.

When simple differences are of interest, a way to partially overcome the dilemma is to say that all possible differences will be of interest and to employ methods appropriate for answering question 1, "Are there any differences in the μ_i, and if so, where are the differences?" There are several multiple comparison procedures available for handling the situation, but we shall discuss only one, the protected LSD procedure. The procedure is implemented in two steps. With Step 1, the first part of question 1, "Are there any differences in the μ_i?" can be answered by an F-test. Let $\bar{x} = \sum_i \bar{x}_i/k = \sum_i[\sum_j x_{ij}/n]/k$ and compute

$$s_b^2 = \frac{1}{k-1}[(n\ \Sigma\bar{x}_i^2 - k\bar{x}^2)] \ .$$

Then $F_{cal} = s_b^2/s_0^2$ is distributed as an $F(k-1,\ k(n-1))$ random variable if $\mu_1 = \mu_2 = \dots = \mu_n$, and should be larger than $F(k-1,\ k(n-1))$ otherwise. If we deem F_{cal} to be too large to be a reasonable observation from the distribution of $F(k-1,\ k(n-1))$, then we say that there is evidence that there are differences in the μ_i. It is important to note that an F-test can only tell us whether there may be differences in the means μ_i, but that it gives no indication of any kind, as to where the differences may be. This must be determined by Step 2 of the protected LSD procedure, which is a series of t-tests. Let some constant significance level α be chosen for the individual t-tests; then a simple method of performing the tests is to compute the α-level least significant difference, $LSD_\alpha = t_0\sqrt{2/n}\ s_0$, where t_0 is the value of $t(k(n-1))$ for which $Pr(-t_0 < t(k(n-1)) < t_0) = 1-\alpha$. The individual differences $\bar{x}_i - \bar{x}_j$ are then compared directly with LSD_α, and any differences exceeding LSD_α are said to be significant at level α.

Example 20: Suppose the following experiment was performed. To determine if four particular limestone types showed any detectible differences in reactivity under laboratory conditions, six 10-gram samples of each limestone type, pulverized to a Fisher size index of 3.0, were subjected for 1 hour at 2400°F to a simulated flue gas mixture having an initial SO_2 concentration of 2000 ppm. At the end of each test, each sample was analyzed for per cent conversion by determining the fraction of unreacted limestone. The data summary and calculations are given as follows:

Limestone Type:	A	B	C	D
	54	68	65	45
	62	81	83	56
% Conversion:	58	87	79	39
	67	72	61	54
	46	75	53	60
	85	67	66	58
Means:	62	75	66	52

$s_b^2 = 545.3$ with 3 degrees of freedom

$s_0^2 = 100.9$ with 20 degrees of freedom

$F_{cal} = 5.4$ with 3 and 20 degrees of freedom

$LSD_{0.05} = \sqrt{100.9/3} \times 2.086 = 12.1$

Step 1. The probability of an F(3, 20) as large as 5.4 is less than (about 0.0075), so there is strong evidence of some differences somewhere.
Step 2. A convenient way to carry out the procedure is to use a ranking and underlining method, whereby series of underlinings are made under the ranked means with the interpretation that any group of means with a continuous line under them are to be considered not significantly different from one another. It may also be helpful to rank the means by simply plotting them on a horizontal axis. Figure 4, on the next page, illustrates Step 2.

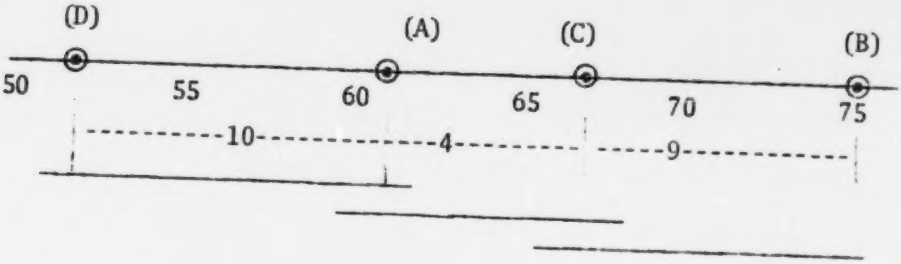


Figure 4. Results of the 0.05-level LSD test ($LSD_{0.05} = 12.1$).

Thus it is concluded that at the 0.05-level, the significant differences are $\mu_C - \mu_D$, $\mu_B - \mu_D$, and $\mu_B - \mu_A$.

Let us now turn to question 2, "Can some grouping pattern in the μ_i be determined?" Unfortunately, although this question is frequently asked as a result of experiments involving several (possibly) different populations, very little has been done toward finding viable procedures to answer it. The majority of existing multiple comparison procedures do not yield solutions, some because of logical complications. Consider, for example, what happens when we try to use the protected LSD procedure to determine a grouping pattern. On the basis of the results of the 0,05-level LSD, as seen above, Limestones D and A belong together, Limestones A and C belong together, and Limestones C and B belong together. Logically, then, all four belong together. But that cannot be, because Limestones D and C are different, as are Limestones D and B, as well as Limestones A and B. Thus, a logical contradiction is reached.

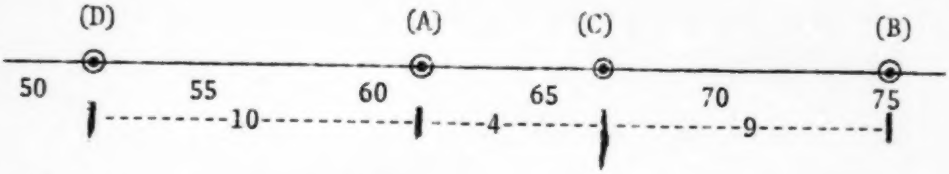
Recently, two procedures have been proposed by John R. Murphy (28), which have been designed specifically as grouping procedures. They are the studentized maximum gap procedure and the studentized range-maximum gap procedure. Let us discuss the second procedure. For a sample $(x_1, x_2, \ldots, x_n)$ from a normal (μ, σ^2) distribution, let $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ represent the ordered sample; i.e.,

$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}, \ldots, x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$. Then, over repeated sampling, the statistic $(x_{(n)} - x_{(1)})/s$ is called the *studentized range* (with n-1 degrees of freedom) and has a well-known distribution which has been computed and tabulated. (Some of the multiple comparison procedures not discussed are, in fact, based on the distribution of the studentized range.) The appropriate studentized range statistic for a set of k means $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$, each based on n observations, is $\sqrt{n}(x_{\max} - x_{\min})/s_0$ with $k(n-1)$ degrees of freedom. With a table of critical values of the studentized range available, the studentized range-maximum gap procedure is carried out as follows:

1. Rank the k means (order of increasing magnitude, say) and compute the magnitude of the gaps (differences between adjacent ranked means).
2. Compute the studentized range for the whole group and test for significance.
3. If the studentized range is judged not significant, stop testing. If the studentized range is judged significant, break the means into two groups by separating it between the two means having the largest gap between them.
4. Repeat steps 2 and 3 for the two new groups formed and continue repeating 2 and 3 for each of the new groups at the previous stage until no more studentized ranges are declared significant.

The grouping of the means is then determined by the breaks declared over all stages.

Example 21: For the limestone data of the previous example, we calculate $s_0/ n = 100.9/6 = 4.1$ with 20 degrees of freedom. The ranked and plotted means were:



The observed studentized range for all four means is $23/4.1 = 5.6$ with 20 degrees of freedom, and the probability of a value this large is less than 0.01 . Thus, a break is declared between $\bar{x}_D$ and $\bar{x}_A$, and at this stage, we now have two groups, D by itself, and A, C, and B together. To determine whether the group of three should be further broken down, we calculate the studentized range for the three means as $13/4.1 = 3.17$. The probability of a value as large as 3.17 is about 0.09 . One might or might not call this a significant result, depending upon his inclination; but if 3.17 is judged to be a significant value for the three-mean studentized range, then another break would be declared between $\bar{x}_C$ and $\bar{x}_B$. The difference between $\bar{x}_A$ and $\bar{x}_C$ would not be significant, and the final grouping would then consist of three groups: Limestone D by itself, Limestones A and C together, and Limestone B by itself.

5.4.5 Analysis of Variance

Inferences about the means of several independent normal populations with common variance are sometimes presented within the context of a one-way classification analysis of variance. The analysis of variance is a very useful statistical technique and is much more general than what will be presented here. Its basis derives from the fact that for any sample or group of samples, the total sample variance can be algebraically partitioned into components which are usually meaningful. Let us refer to the samples from the different normal populations of the previous section as classes, and show how the partitioning of the total sample variance may be done for that case. Recall that we had a total of kn observations $(x_{11}, x_{12}, \ldots, x_{1n}), \ldots, (x_{k1}, x_{k2}, \ldots, x_{kn})$, where it was assumed that for each $i = 1, 2, \ldots, k$ $(x_{i1}, x_{i2}, \ldots, x_{in})$ was a sample from a normal (μ_i, σ^2) population, with some of the μ_i being, perhaps, the same. Now, each observation x_{ij} may be represented as,

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i),$$

where $\bar{x}_i = \sum_j x_{ij}/n$ and $\bar{x} = \sum_i \bar{x}_i/k = \sum_{ij}\sum x_{ij}/kn$. Thus,

$$(x_{ij}-\bar{x})^2 = (x_i-\bar{x})^2 + 2(\bar{x}_i-\bar{x})(x_{ij}-\bar{x}_i) + (x_{ij}-\bar{x}_i)^2 .$$

For fixed i, sum both sides with respect to j. Then

$$\begin{aligned}\sum_j (x_{ij}-\bar{x})^2 &= \sum_j (\bar{x}_i-\bar{x})^2 + 2(\bar{x}_i-\bar{x}) \sum_j (x_{ij}-\bar{x}_i) + \sum_j (x_{ij}-\bar{x}_i)^2 \\ &= n(\bar{x}_i-\bar{x})^2 + \sum_j (x_{ij}-\bar{x}_i)^2.\end{aligned}$$

The cross-product term vanishes, because $\sum_j (x_{ij}-\bar{x}_i) = \sum_j x_{ij} - n\bar{x}_i = 0.$ Now, summing on i,

$$\sum_{ij}\sum (x_{ij}-\bar{x})^2 = n\sum_i (\bar{x}_i-\bar{x})^2 + \sum_{ij}\sum (x_{ij}-\bar{x}_i)^2 .$$

The quantity on the left is called the total sum of squares, and the two quantities on the right are called, respectively, the sum of squares between classes and the sum of squares within classes. The sum of squares between classes measures the failure of the class means to all be the same value $\bar{x}$; that is, their variation about $\bar{x}$. For fixed i,

$\sum_j (x_{ij}-x_i)^2$ measures the failure of the observations in the ith class to be equal to the class mean $\bar{x}_i$; that is, the variation of the observations in the ith class about the ith class mean $\bar{x}_i$. Adding together all k of the within-class variation terms, we get the total within-class variation term, the sum of squares within classes. Now, note that

$$s_b^2 = \frac{1}{k-1}\left[n\left(\sum_i \bar{x}_i^2 - k\bar{x}^2\right)\right] = \frac{n}{k-1} \sum_i (\bar{x}_i-\bar{x})^2 .$$

Also, recall that

$$\begin{aligned} s_0^2 &= (s_1^2 + s_2^2 + \dots + s_k^2)/k \\ &= \frac{1}{k(n-1)}[(n-1)s_1^2 + (n-1)s_2^2 + \dots + (n-1)s_k^2] \\ &= \frac{1}{k(n-1)}[\Sigma(x_{1j}-\bar{x}_1)^2 + \Sigma(x_{2j}-\bar{x}_2)^2 + \dots + \Sigma(x_{kj}-x_k)^2] \\ &= \frac{1}{k(n-1)} \sum_{ij}\sum (x_{ij}-\bar{x}_i)^2 . \end{aligned}$$

It is now evident that the sum of squares between classes and the sum of squares within classes are just s_b^2 and s_0^2, multiplied by their respective degrees of freedom, k-1 and k(n-1). The information can be neatly summarized in an analysis of variance table, sometimes referred to as an AOV or ANOVA.

ANALYSIS OF VARIANCE TABLE FOR A ONE-WAY CLASSIFICATION

WITH EQUAL CLASS SIZES

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{cal}
Total	nk - 1	$\sum_{ij}\sum(x_{ij}-\bar{x})^2$	-------	
Between	k - 1	$n\sum_i(\bar{x}_i-\bar{x})^2$	$\frac{n}{k-1}\sum_i(x_i-\bar{x})^2$	$F = \frac{\text{between mean sq}}{\text{within mean sq}}$
Within	k(n - 1)	$\sum_{ij}\sum(x_{ij}-\bar{x}_i)^2$	$\frac{1}{k(n-1)}\sum\sum(x_{ij}-\bar{x}_i)^2$	(with k-1 and k(n-1) degrees of freedom)

A feature of an AOV table is that the degrees of freedom and sums of squares columns "add up;" for the table above, in both columns, Total = Between + Within. This property makes an AOV table a convenient means of "variation accounting," and is partly responsible for its great appeal. Another appealing feature is that one or more of the mean square quantities (which are simply the corresponding sums of squares, divided by their respective degrees of freedom) can often be regarded as estimates of things of interest. For example the "Within Mean Square" above is actually s_0^2, the best unbiased estimate of the common unknown variance σ^2 of the normal populations from which the k samples were assumed to have been taken. In some complex cases, the AOV mean squares are the only way to obtain reasonable estimates of unknown population variance.

Let us construct an AOV table for the Limestone data of the previous section. As a first step, the expressions which define the sums of squares do not lend themselves well to computation, and can be put into more convenient forms for computation purposes:

1. Define $x_{..} = \sum_{ij}\sum x_{ij}$ and $x_{i.} = \sum_j x_{ij}$, so that $x_{..}$ is the total of all observations and $x_{i.}$ is the total of the observations in the ith class.
2. Since $nk\bar{x} = \sum_{ij}\sum x_{ij}$, define the "correction factor" as

   $$CF = nk\bar{x}^2 = (\sum_{ij}\sum x_{ij})^2/kn = x_{..}^2/kn\ .$$

   Then

   $$\text{Total Sum of Squares} = \sum_{ij}\sum(x_{ij}-\bar{x})^2 = \sum_{ij}\sum x_{ij}^2 - nk\bar{x}^2 = \sum_{ij}\sum x_{ij}^2 - CF.$$

3. Since $n\bar{x}_i = \sum_j x_{ij} = x_{i.}$, $n\bar{x}_i^2 = x_{i.}^2/n$. Thus,

$$\text{Between Sum of Squares} = n\sum_i(\bar{x}_i-\bar{x})^2 = n\sum_i\bar{x}_i^2 - nk\bar{x}^2$$

$$= \sum_i x_{i.}^2/n - CF\ .$$

4. $$\text{Within Sum of Squares} = \sum_i[\sum_j(x_{ij}-\bar{x}_i)^2] = \sum_i(\Sigma x_{ij}^2-n\bar{x}_i^2)$$

$$= \sum_i(\sum_j x_{ij}-x_{i.}^2/n)$$

Now, for the Limestone data, $n = 6$, $k = 4$, $\sum_{ij}\sum x_{ij} = x_{..} = 1530$,

$\sum_{ij}\sum x_{ij}^2 = (54)^2 + (62)^2 + \ldots + (60)^2 + (58)^2 = 101192$, and $CF = (1530)^2/24$

$= 97537.5$. Thus

$$\text{Total Sum of Squares} = 101192 - 97537.5 = 3654.5$$

$$\text{Between Sum of Squares} = [(372)^2 + (450)^2 + (396)^2 + (312)^2]/6 - 97537.5$$

$$= 99174 - 97537.5 = 1636.5$$

$$\begin{aligned}\text{Within Sum of Squares} = \ & [(54)^2 + \ldots + (85)^2 - (372)^2/6] \\ + \ & [(68)^2 + \ldots + (67)^2 - (450)^2/6] \\ + \ & [(65)^2 + \ldots + (66)^2 - (396)^2/6] \\ + \ & [(45)^2 + \ldots + (58)^2 - (312)^2/6] \\ = \ & 890 + 302 + 488 + 338 = 2018\end{aligned}$$

Note that the Within Sum of Squares could have been obtained by subtracting the Between Sum of Squares from the Total Sum of Squares, and this is usually the recommended procedure; however, computing it both ways does provide a check on the arithmetic. The AOV table is, then,

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{cal}
Total	23	3654.5	-	
Between Limestones	3	1636.5	545.5	5.41
Within Limestones	20	2018.0	100.9	

As a final word, let us sound a note of caution. The analysis of variance has become a popular statistical tool to data analysts everywhere, and rightly so, for it is convenient and orderly, and sometimes it is the only way to derive certain estimates. There is a tendency for some, however, after gaining some familiarity with the techniques involved, to perhaps regard analysis of variance and analysis of data as one and the same thing. This, of course, is a mistake, in that the analysis of variance is but one of several statistical tools available for the data analyst to draw upon.

6. EXPERIMENTAL TEST PROGRAMS

6.1 Introduction

In recent times it has become more and more common to see considerations of "statistical design" cropping up in experimental testing programs, and it is reasonable to ask "why?". Perhaps, as some contend, it is just a faddish thing which, in a few years, will cease to be fashionable; but, as ones who are called upon to apply the principles of statistical design time and again, we prefer to believe otherwise. In that regard, we should note that, as with most disciplines, statistics is a dynamic subject, constantly changing to address new problems and to deal more effectively with old ones and, as a result, specific techniques will come and go. However, the principles upon which the various techniques are based will cease to be valid only when our present methods of scientific reasoning become absolute. Thus, our answer to the question of why the prevalence of statistical design in experimentation is that the underlying principles have been found to be useful and applicable in a wide variety of experimental situations.

In this short chapter on experimental design we hope to accomplish the following: (1) introduce the reader to some of the basic concepts of experimental design, (2) equip the reader with enough of the basic ideas so that he can make some application of the subject of experimental design, and (3) help the reader to consult with a statistician on the problems of experimental design.

As far as the first objective is concerned, it should not be expected that a single course or a single book on any subject will provide an in-depth knowledge of a subject. It can only provide an introduction. To the practicing engineer who has designed many tests and experimental programs, the need for an introduction may not appear obvious. We can only say that many scientists and engineers have benefitted from an introduction to experimental design.

We should hope that, in addition to providing an introduction, this effort will also provide enough information so that the readers for whom it is intended will actually be able to make use of it in their work. It will shortly be apparent that most of the material of this chapter can be studied independently of the first four chapters and most of the fifth chapter. Therefore, it is possible for the reader to jump very quickly to what we have to say about experimental design. It apparently was the purpose of Snedecor in his monumental work (32) to provide the researcher with methods which he could apply almost immediately. We hope to emulate his example. Of course it is not our purpose to write a textbook, giving a thorough coverage of every topic.

Despite the natural human inclination to want to do everything for ourselves, it is often worthwhile to avail ourselves of the expertise of a consultant. Untold numbers of scientists and researchers have found their work strengthened and clarified by consulting with a statistician in the planning stages of their research. The researcher going through this experience for the first time may find that the statistician uses many seemingly strange terms and unfamiliar concepts. It is hoped that this chapter will help the reader to benefit from the experience of consulting with a statistician.

6.2 Data

In the scientific disciplines, we are accustomed to thinking of data as information arising from purposefully designed investigations. However, we should remind ourselves that, as the term is commonly used, it refers to information arising from any source whatever. We hear a lot nowadays about the "data explosion," and it therefore seems appropriate to discuss briefly some interrelated aspects of the subject of data. First, let us consider how it may be viewed from the standpoint of answering the question, "What should be done with it (or to it)?" Earlier, it was pointed out that the function of condensing and summarizing the information in a set of data is not generally included within the category of statistical inference. However, in view of recent technical improvements in our ability to obtain more data and obtain it faster, the problem should by no means be considered trivial or unimportant. There is no doubt that more imaginative and efficient methods of extracting and summarizing salient information from data are needed, and some of the tools of statistics can be of great assistance in that regard. However, emphasis upon the ability to obtain more and more data sometimes goes hand-in-hand with the attitude that the more data, the better, and it is this attitude we wish to criticize. Such thinking is to be deplored because it fosters the belief that if a large enough data base is created, then the probability of finding a meaningful trend somewhere is greater. It can also lead to the belief that it is valid to search a set of data until some "unusual" trend is discovered, and then to proceed as if the subsequent conclusions based on this trend were fact, rather than hypotheses to be further examined. We regard such behavior as totally unscientific.

Let us now examine the gathering and analysis of data in the context of the scientific method. While there are several important facets, let us agree that the scientific method is a cyclic activity with two important but separate parts to the cycle. One part of the cycle is the study of existing data for trends, and combining the resulting discoveries with established knowledge to form tentative hypotheses about those trends. The other equally important part of the cycle is the purposeful experimentation (gathering of new data) designed for the purpose of determining whether the hypotheses are supported or discredited by further observation. It is evident that the practices described in the preceding paragraph emphasize the first part of the cycle to the almost total exclusion of the second. Thus, what it comes down to is that data should be taken with a purpose in mind. One must have a clear idea of what a set of data is to show in order to decide after the fact whether this or that trend has been demonstrated. A large data base of "historical" information is valuable only when it leads to theories which may be tested or questions which may be answered by further experimentation. The gathering of new data is justifiable only when it serves to resolve the questions raised. The cyclic nature of the process is evident, because the new data becomes historical data, and frequently suggests further hypotheses or questions to be resolved, and the cycle is thus repeated.

For all the remaining discussion, we shall be talking about gathering and analysis of data with respect to the second part of the investigative cycle; that is, data which is collected as the result of a purposeful experiment. It can be argued that all data contains useful information, but let us remind ourselves that we are dealing in terms of statistical and scientific inference. In such a case, we must be able to consider the data under examination to be a legitimate and useful sample from some larger population which we wish to say something about. Usually it is difficult, if not impossible, to justify any relation between sample and population for some particular set of data when the purpose and manner of its taking are not known. Even when both of these requirements are met, it is sometimes an eye-opening exercise to stop and consider what the legitimate population of inference really is. One can be quite surprised and thoroughly disappointed to find how restricted that population is compared to the population he really wanted to sample. Thus, the advantage of carefully planned data collection is that we are thereby forced to think about why and how it is to be done which, in turn, requires that we define the population of inference. It is only then that the principles described throughout this handbook can be applied. To reiterate, once that population has been delineated, then one can do his best to draw a random sample from it. And we have seen that, as a consequence of such a procedure, the powerful mathematics of the theory of probability may be applied to the problem, which can give us a measure of the risk and uncertainty associated with the inferences that are drawn from sample to population. The upshot of the preceding discussion is that the analysis and interpretation of data can be accomplished with at least a degree of objectivity, rather than being based largely upon opinion and subjective judgement, and that such an accomplishment is made possible by proper data gathering techniques.

To summarize, in this section, we have pointed out that the term "data," in relation to inference, means something other than a volume of information to be sorted through and manipulated. In a manner of speaking, we are actually referring to information which has not yet been collected. We have once again emphasized the importance of the data gathering step and have stressed how careful planning can reduce the difficulties and risk associated with drawing inferences. However, the heart of the matter is this: regardless of whether careful planning is involved or not, the manner in which the data is collected must be known, for the analysis and interpretation (inferences to be drawn) depend directly upon it.

6.3 Comparative Experiments

A definition of "experiment" given by Webster's dictionary is "A test made to prove or disprove something doubtful; an operation undertaken to discover some unknown principle or to test some known truth; as, a laboratory experiment in physics."

Although there is little difficulty in achieving a consensus on the meaning of the words experiment or experimental, it seems that there is greater difficulty in attempts to distinguish between two types of experiments, comparative and absolute. This distinction is discussed by Kempthorne (22) who references Anscombe (1).

Basically the idea of an absolute experiment is that the experiment is being conducted to determine the absolute value of some quantity; for example, the speed of light. The world of physics and chemistry abounds with physical constants, usually identified with some particular person's name which resulted from absolute experiments. For example, we mention Avogadro's constant, Faraday's constant, and Planck's constant.

Comparative experiments on the other hand are concerned with measuring the differences in the effects of two or more stimuli (called treatments) upon some characteristic (or characteristics) of interest. These differences are almost always measured in an absolute sense. This has sometimes seemed strange to engineers more accustomed to thinking of relative differences or percentage changes. However, almost all theory and methodology in the design of experiments deals with absolute differences.

6.4 The Use of the Word "Design"

Although there are subsets of the statistics community for which the term "experimental design" has a clear and agreed-upon meaning, the expression is used in a bewildering variety of ways by statisticians (for example those listed in the 1973 directory of statisticians published by the American Statistical Association). Of course the practice of statistics is not confined to this restricted group of people. There are other large groups of people with vital interests in statistics and specifically in the field of experimental design. The result is that a novice may have great difficulty in comprehending the way in which the word "design" is being used. In order to discuss the matter further, let us refer to Figure 5.

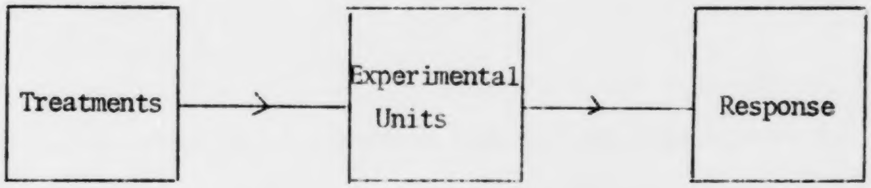


Figure 5. Experimental design .

In a comparative experiment, certain stimuli, called treatments, are applied to experimental material, producing responses. If the responses are not already quantitative, they are converted to quantitative responses. An analysis is then performed to compare the different stimuli with respect to the quantitative responses. All of this should be done in a manner following existing theory and methodology rather than in a completely ad hoc pattern. The word "design" seems like a perfectly suitable word to describe the plan for conducting and analyzing the experiment. The difficulty with the word "design" arises because it is used to describe (among other things): (1) the treatments, (2) the experimental material, (3) the way in which treatments are assigned to the experimental material, (4) the responses, and (5) the analysis. *In this handbook, we use the term "experimental design" to describe the experimental material and the way in which treatments are assigned to the experimental material*. This is consistent with well-known books on experimental design, Cochran and Cox (8) and Kempthorne (22). Thus, we use "experimental design" to include topics (2) and (3), above. The use of the word "design" to describe topics (1), (4), and (5) will be discussed in later sections.

6.5. Properties of a Good Experiment

In an excellent chapter on comparative experiments, D.R. Cox (10) lists the following requirements for a good experiment: (1) absence of systematic error, (2) precision, (3) range of validity, (4) simplicity, and (5) the calculation of uncertainty. We will now illustrate these five requirements as we understand and interpret them.

First, an experiment should be free from systematic error. Although this statement seems totally obvious and compelling, it is sometimes flagrantly violated either knowingly or not. As an example, suppose that an experiment was designed to compare particulate mass emission rates resulting from the use of two gasolines in test automobiles. Although the test equipment was quite sophisticated, there were only two automobiles used in the test. Automobile A used the first gasoline and automobile B used the second gasoline. Regardless of laboratory techniques and precision of measurement, the comparison of gasolines is handicapped because of the systematic error incorporated. Any comparison of gasolines is also a comparison of automobiles. Needless to say, such systematic errors should be avoided if at all possible. The primary technique which is used to eliminate systematic errors is that of randomization. This will be discussed in some detail a little later.

Secondly, let us discuss the matter of precision in a comparative experiment. For simplicity, suppose we have two treatments. Inevitably, some statistic will be calculated as representing the difference in the effects of these two treatments. In the simplest cases this statistic will be simply the difference of two sample means. If we say that the experiment has high precision, we mean that the variance of this difference is small, the variance being defined for the conceptual population of differences which would result from repetition of the experiment. For example, suppose that an experiment was designed to compare the outlet dust concentration (gr /10^3 ft^3) in two types of filter fiber material, cotton and nylon. The experimental results give a difference of 105. The importance of this result depends heavily upon its precision. If it were known that the inherent standard deviation is 50, say, the results would be interpreted much more skeptically than if it were known that the standard deviation is 25, for example.

While the need for precision may be clearly perceived, it is not so obvious how it is to be achieved. The variance and precision of such differences as discussed above depend upon: (1) the inherent variability of the experimental material, (2) the number of observations, and (3) the design of the experiment. While the experimenter can do nothing about the first factor, he can exercise considerable control over precision through the second and third factors.

One other comment needs to be made about precision. The term is widely used in statistics to discuss the amount of variation rather than the average or central value. It is not to be confused with accuracy which is used to discuss the amount of bias in a result.

The third property of a good experiment alluded to above was range of validity. The basic idea is that the experiment should encompass a sufficiently wide range of treatments experimental material, etc., so that the inferences drawn from the experiment may with some validity be made about the population of interest. All of us are aware that laboratory experiments on SO_2 emission may not apply at all to industrial SO_2 emissions over Birmingham, for example. Yet, this point must be emphasized over and over again. It is not possible in any statistical sense to make inferences about populations not sampled.

The desire to have a wide range of validity is at direct odds with the motivation to conduct a precise experiment. Incorporating a wide range of conditions into an experiment tends to decrease the precision of the experiment. On the other hand, carefully controlled laboratory experiments with great precision may have such a narrow range of validity that the results are virtually of no interest.

The fourth requirement of a good experiment is that it be simple. This does not mean that it should be naive or stupid. On the other hand, experiments designed by statistical novices are often too ambitious and much too complicated, and the experiment may be doomed to failure before it is ever begun. An experiment should be complex enough that it utilizes available theory and methods but simple enough that it can be conducted and analyzed easily.

The fifth requirement listed by Cox was the calculation of uncertainty. We interpret this requirement as meaning that the experiment should be performed in such a way that the analysis can be based upon theoretical development which has been field-tested. Generally, the theory for any analysis will begin with a probability model; hence the expression "calculation of uncertainty." It is a curious situation that well-trained people will often embark upon data collection and analysis armed only with ad hoc procedures, inventing statistical methods as they go, even though they would not consider following the same practice with respect to chemical or physical theory or laboratory methods.

6.6 Experimental Units

The concept of experimental unit is a central one in the theory of experimental design but is apparently not an easy concept to comprehend. Stated simply, we define an experimental unit as the smallest subdivision of experimental material so that different subdivisions of the material have the opportunity to receive different treatments. The opportunity to receive different treatments refers to the opportunity in the population of conceptual repetitions. Obviously, this depends to some extent upon the scheme used for assigning treatments to the experimental material and therefore upon the randomization. Thus, a discussion of experimental units is bound to that of randomization, but we shall attempt to simplify the discussion by separating the two topics.

Consider the following experiment to determine the rate of sulfate ion formation in water droplets. Droplets of water were placed on a grid of Pyrex fibers and suspended in a chamber with a known concentration of SO_2. The amount of H_2SO_4 formed during the reaction period was then calculated. The laboratory technicians were reluctant to make many changes in the SO_2 concentration so the experiment was done as follows. The SO_2 level was fixed and four grids were run through the experiment. The level was changed and four more grids were run. The level was changed a third time and four more grids were run. Although there are 12 determinations of H_2SO_4 , there are really only three experimental units. Note that the first four grids do not have an opportunity to receive different levels of SO_2.

As a second example illustrating the idea of experimental units, consider an experiment to study the usefulness of plants for the monitoring of air pollutants. Two test chambers were used. Twelve plants were placed in test chamber A which was filled with unfiltered air. Twelve plants were placed in test chamber B which was filled with filtered air. After a specified period of time, the fluorine content (mg/100g dry weight) was measured for each of the plants. Although the experiment involved 24 plants, there were really only two experimental units. The experiment could have been modified to involve 24 experimental units by doing the following: using only one plant per experimental chamber, refilling the chambers after each run, and randomizing the order so that the sequence might appear as U,F,F, U... etc., where F stands for filtered and U stands for unfiltered.

6.7 Experimental Error and Sampling Error

Experimental error is as important to the subject of experimental design as experimental unit. To understand the concept of experimental error let us return to the first example in the previous section, that of H_2SO_4 formation. Having completed the experiment, consider the average responses between levels A and B of SO_2, x_A-x_B. When we calculate the difference $\bar{x}_A-\bar{x}_B$, it can be written as

$$x_A - x_B = \mu_A-\mu_B + e$$

where $\mu_A-\mu_B$ represents the difference in population means of the two concentrations of SO_2 and e represents a random error. The population referred to is the conceptual population which would result from repetition of the experiment. The random error e is composed of all those variables which cause the difference of sample means to be in error and is called the experimental error. It is extremely worthwhile to attempt an enumeration of these variables.

The variables which contribute to the experimental error are

1. Differences between experimental units.
2. Errors in the application of treatments.
3. Measurement errors.
4. Other factors either known or unknown but neglected in the design of the experiment.

The experimental error e is commonly assumed to have zero mean and variance σ^2. The assumption of zero mean implies that we have designed our experiment so that we have an unbiased estimate of $\mu_A-\mu_B$.

Quite frequently, the experimental error variance is simply referred to as experimental error. This may cause some confusion but need cause no concern to the person consulting the statistician. It is simply a matter of learning the terminology. Interest lies in estimating the variance of e, the experimental error variance, so that we can express our uncertainty about $\mu_A - \mu_B$. The emphasis has been almost totally on unbiased estimation of σ^2.

One of the most bewildering experiences for nonstatisticians in consulting with statisticians occurs when the estimation of experimental error is discussed. To the statistician's statement, "there is no experimental error," or "there are no degrees of freedom left for error," or "you have no estimate of experimental error," the non-statistician either nods in agreement or leaves in silent resentment. We would like to give an introduction to the statistician's thought processes about estimation of experimental error variance.

The matter almost always arises in analysis of variance situations. In the simplest case, the experimental error variance σ^2 is estimated by a quantity which is equivalent to a constant times a sum of squares of deviations about the mean; i.e.

$$c \, \Sigma(x_i - \bar{x})^2.$$

The problem is in deciding which sum of squares is appropriate for estimating the experimental error variance. Sometimes the highly trained mathematical statistician despairs of trying to understand how the applied statistician makes this decision. This decision making is partially an art, partially a science; of course, a thread of rigor runs through it. There is no way that the novice can easily master this idea but we believe that we can introduce a few of the important ideas.

The question of whether or not a particular sum of squares is appropriate for estimating σ^2 can be studied by the use of the following identity

$$\Sigma(x_i - \bar{x})^2 = \frac{1}{2n} \sum_{i<j} (x_i - x_j)^2.$$

The sum of squares of deviations about the mean can be calculated from squares of difference between individual numbers. The basic principle in deciding whether or not a given sum of squares is appropriate is that every difference $x_i - x_j$ should contain all of the variables that enter into the experimental error e.

Let us now illustrate the concept with the last example in Section 6.6 . Recall that the example called for putting 12 plants in a chamber with filtered air and 12 plants in a second chamber with unfiltered air. Denote the responses as follows

First Chamber	Second Chamber
x_{11}	x_{21}
x_{12}	x_{22}
x_{13}	x_{23}
x_{14}	x_{24}
x_{15}	x_{25}
x_{16}	x_{26}
x_{17}	x_{27}
x_{18}	x_{28}
x_{19}	x_{29}
$x_{1,10}$	$x_{2,10}$
$x_{1,11}$	$x_{2,11}$
$x_{1,12}$	$x_{2,12}$

The obvious procedure is to calculate the difference of means

$$\bar{x}_1 - \bar{x}_2$$

and variance

$$\sigma^2\left(\frac{1}{12} + \frac{1}{12}\right) .$$

Now, how should σ^2 be estimated. The obvious procedure (though incorrect) would be to estimate σ^2 by

$$\frac{\Sigma(x_{1j}-\bar{x}_1)^2 + \sum_j(x_{2j}-\bar{x}_2)^2}{11 + 11} .$$

To see that this is not a proper estimate of σ^2, consider the differences entering into the sum of squares in the numerator. None of these differences contain the difference between chambers. The difference between chambers, after all, is likely to be one of the major variables which cause $\bar{x}_1 - \bar{x}_2$ to be in error.

The estimate of σ^2, just considered, is an unbiased estimate of another variance, called the variance of sampling error. Sampling error simply contains all of the variables which cause observations resulting from the same treatment to be different, and generally has a much smaller variance than experimental error. This point is often not appreciated with the result that the experimental error variance is badly underestimated and, to make matters worse, too many degrees of freedom are assigned to experimental error.

6.8 Degrees of Freedom in the Analysis of Variance

We now want to try to describe a phenomenon which is not well described in any text book-- probably for good reason. It is difficult. Nevertheless, the user of statistics will soon become aware of the ability of applied statisticians to think and work in terms of degrees of freedom in the analysis of variance.

For over fifty years (See Fisher and MacKenzie (16)) the analysis of variance has been a powerful tool not only for analyzing experimental data, but also for thinking about the design of the experiment. Basically the analysis of variance is a partition of the sum of squares of deviations $\Sigma(x-\bar{x})^2$ into recognizable component **sum of squares**, each with its own degrees of freedom. If there are n observations in the experiment, there are (n-1) degrees of freedom associated with the total sum of squares $\Sigma(x-\bar{x})^2$. It is always possible to write this as the sum of n-1 component sum of squares, each with 1 degree of freedom. Usually these component sum of squares would not be interpretable or recognizable. Instead, there will generally be fewer than n-1 component sum of squares, each with more than 1 degree of freedom.

The experienced statistician develops the ability to describe the appropriate analysis of variance and the breakdown of degrees of freedom. To do this he leans heavily upon the concepts of experimental unit, randomization, and estimation of experimental error variance.

6.9 Randomization

The idea of arranging a set of numbered objects in random order is another of the central concepts in experimental design. This can be accomplished for small sets of objects by the use of physical processes like drawing cards from a deck, marbles from a bowl, or tickets from a container. The primary methods of randomization which are used, however, are either the use of random number tables or computerized random number generators. In both cases, it should be realized that we are simulating a random device.

There are many different random number tables available. These tables present digits zero through 9 hopefully "scrambled" together in something resembling random order. Hald (18) presents numbers compiled from drawings in the Danish State Interest Lottery of 1948. Fisher and Yates (17) present a table of digits abstracted form Logarithmetica

Brittanica consisting of the 15th to 19th digits of logarithms to the base 10. The Rand Corporation (30) presents a table of a million random digits generated from a computer program.

Numerous computer programs exist to generate sequences of random digits. Many of these generating routines are refinements of the following simple procedure described by Kempthorne and Folks (23):

1. Choose a multiplicand.
2. Multiply by an arbitrary multiplier.
3. Use the k digits on the right for the second multiplier and record the p digits immediately preceding these k.
4. Multiply the multiplicand by the second multiplier, and continue the process.

Using a constant multiplicand of 341 and an initial multiplier of 443 we obtain a product of 151,063. Recording the fourth digit from the right, "1", and using the three digits on the right, 063, as a new multiplier, we obtain a second product of 21, 483. Recording "1" and using 483 as the third multiplier we proceed. This results in the following sequence of 100 digits.

1	9	7	4	2	1	9	7	4	2
1	9	7	6	4	1	9	7	6	4
4	2	1	8	6	4	2	1	8	6
9	7	5	3	1	9	7	5	3	1
6	4	1	0	8	6	4	1	0	8
5	3	1	8	6	5	3	1	8	6
5	3	1	0	8	5	3	1	0	8
8	7	5	2	0	8	7	5	2	0
3	1	9	7	5	3	1	9	7	5
0	7	6	4	2	0	7	6	4	2

If continued, the process will simply repeat this sequence of 100 digits. This periodicity is typical of such generating schemes. Naturally, in order to be useful, the period should be quite large. For the sequence we have generated, the frequency of digits is as follows:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	8	16	8	8	10	10	10	12	10	8

These frequencies appear to be reasonable outcomes for a random device which would generate each digit with equal frequency.

Published random number tables have been scrutinized not only with respect to the frequency of occurrence of digits, but with respect to runs of the same digit, gaps between successive occurrences of the same digit, etc. Excessive irregularities have been eliminated. The question often arises about how we can call any table of digits "random" particularly after it has been "doctored." Remember that we want the use of the table to simulate a random number generator. Thus we want the table to be reasonably free from irregularities so that repeated use of the table might at least resemble repeated use of a random mechanism.

The use of randomization for the assignment of treatments is credited to R.A. Fisher (15). Although not always practiced by experimenter the principle is widely recognized. The random assignment of treatments to experimental units achieves two important effects: (1) the minimization of systematic effects, and (2) a basis for the assumption of a probability model.

Left to his own devices, an experimenter may be more likely to arrange treatments in a systematic fashion than if he makes formal use of a randomization device (or table of random numbers) to assign treatments. Thus the more or less "haphazard" order introduced by randomization would have desirable effects even if there were to be no analysis based on a probability model.

The second effect requires more sophistication to appreciate. We have referred previously several times to a conceptual population of repetitions of the experiment. The immediate effect of randomization is that the conceptual population is clearly defined and the consequence of having used one of the possible permutations of treatments is that we feel justified in using a probability model.

6.10 Randomized Designs

Recall that in Section 6.4 we used the term "experimental design" to mean the description of experimental units and the assignment of treatments to the experimental units. From this viewpoint an experimental design is a plan for generating the data. It is not a description of the analysis to be performed on the data. In the previous section it was indicated strongly that the only method of assigning treatments would be random. Different designs may be thought of in terms of the restrictions placed upon the random assignment of treatments.

Paired Design

There are n pairs of experimental units and two treatments. Within each pair of experimental units, treatment 1 is assigned at random to one of the units and treatment 2 is assigned to the other.

Two-Group Design

There are n experimental units and two treatments. Treatment 1 is assigned at random to n_1 of the units and treatment 2 to the other n_2 units where $n_1 + n_2 = n$.

Completely Randomized Design

There are n experimental units and k treatments. Treatment 1 is assigned at random to n_1 units, treatment 2 at random to n_2 units, . . . , and treatment k at random to n_k units, where $n_1 + n_2 + \ldots + n_k = n$. Quite obviously, a completely randomized experiment is a generalization of the two-group design.

Randomized Block Design

There are n experimental units grouped (or blocked) into b blocks of t experimental units each and there are t treatments. Within each block the treatments are assigned at random to the experimental units. The paired design is a special case of the randomized block design.

Latin-Square Design

There are $n(=t^2)$ experimental units which have been grouped or classified by rows and columns into a $t \times t$ square and there are t treatments. The treatments are assigned to the experimental units subject to the restriction that the resulting array of treatment numbers has the Latin-square property--namely that each number occurs once and only once in each row and each column.

Consider an experiment to study the reaction of fabrics made from synthetic fibers to air contaminated with nitrogen dioxide, ozone, or sulfur dioxide. We wish to use this setting to illustrate the elementary experimental designs just described.

Suppose that ten pieces of nylon were available and that each piece of fabric is cut in half. The treatments are exposure to light and air, and to light and air containing 0.2 ppm SO_2. For each piece of nylon, one of the treatments is assigned to one half and the other treatment to the other half. After exposure, the relative breaking load is measured for each of the ten pairs of pieces of fabric. This is obviously an example of what we called the paired design.

As a variation of this example suppose that, instead of cutting the ten pieces of nylon in half, we simply assign at random one of the treatments to five of the pieces, and the other treatment to the other five pieces. This would be an example of what we called the two-group design.

In the above experiment, there were three treatments: (1) control, (2) exposure to light and air, and (3) exposure to light and air containing 0.2 ppm SO_2. If each treatment is assigned at random to several pieces of nylon fabric, this is an example of a completely randomized design.

Suppose that we have three pieces of nylon fabric from each of four different manufacturers. For the fabric from each manufacturer, the three treatments are assigned at random to the three pieces of fabric. This illustrates a randomized block design.

As a final illustration in this section, we might have nine pieces of nylon fabric representing three different manufacturers and three different years of manufacture. Then the three treatments are assigned at random subject to the Latin-square restriction.

We have not discussed reasons for using the different designs discussed. These matters are discussed at great length in many statistics textbooks; but in our experience it is not always easy to understand the differences between the designs.

6.11 Multifactor Experiments

The problem of studying several experimental variables in an experimental program is an extremely common problem. However, the idea of simultaneously varying all variables of interest is still not universally known or accepted, years after Fisher (14) introduced the idea. In a landmark paper entitled "The Arrangement of Field Experiments," Fisher put forward the idea of multifactor experiments which he called complex experimentation. Fisher stated, "No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed." Instead of studying one factor at a time, multifactor or factorial experiments study all the factors of the experiment.

In terms of the ideas presented in previous sections, the description of a factorial (or multifactor) experiment is not a design. Rather, the factorial description refers to the treatment structure which may be used with various kinds of experimental designs. We may have a two-factor arrangement of treatments in a completely randomized design or in a randomized block or in a Latin-square, etc.

Korth (24) describes a three-factor factorial experiment designed to investigate irradiated auto exhaust under conditions of continuous mixing. The three factors and their levels were:

1. Initial exhaust concentration
   13 ppm carbon and 35 ppm carbon.
2. Average irradiation time
   85 minutes and 120 minutes.
3. Fuel composition
   14% olefins and 23% olefins.

As far as the treatments of the experiment were concerned, there were eight treatments determined by the eight combinations of three factors, each at two levels.

A number of different response variables were looked at but for one of them, NO_2 formation rate in pphm/min, 17 runs were reported as follows:

Exhaust Level ppm carbon	Fuel Composition, % olefins	Average Irrad. Time, min	NO_2 Formation Rate, pphm/min
13	14	85	1.76, 1.25, 1.50
13	14	120	1.45, 1.67
13	23	85	1.88, 2.22
13	23	120	2.10, 1.95
35	14	85	2.42, 2.90
35	14	120	2.61, 2.36
35	23	85	3.59, 3.80
35	23	120	2.68, 2.74

From looking at this table, one can not decide what experimental design was used and therefore what the appropriate analysis should be. There are many possibilities. We suggest three of them: (1) 2 x 2 x 2 factorial in a completely randomized design with 17 experimental units, (2) 2 x 2 x 2 factorial in a completely randomized design with eight experimental units, and (3) 2 x 2 x 2 factorial in a randomized block design with 17 experimental units grouped into two blocks. We now discuss each of these possibilities in greater detail.

1. Completely Randomized Design with 17 Experimental Units

Unequal sample sizes are awkward for trying to present an elementary account. In order to make this presentation simpler, let us assume that there were two runs with all eight treatments. Let us use (1.76, 1.25) instead of (1.76, 1.25, 1.50). Then we consider the experimental units to be the 16 runs and each of the eight treatments would be assigned at random to two of the experimental units. A random assignment might have been as follows, for example:

Run	Exhaust Level, ppm carbon	Fuel Composition, % elegins	Average Irrad. Time, min
1	13	14	85
2	13	23	85
3	35	23	85
4	35	14	85
5	13	14	120
6	13	23	120
7	13	14	120
8	35	14	85
9	35	14	120
10	35	23	85
11	13	23	120
12	35	14	120
13	35	23	120
14	13	14	85
15	13	23	85
16	35	23	120

The appropriate analysis of variance for this randomization is shown in the following table

AOV (1)

Source	df	SS	MS	F
Total	15	7.563975		
Treatments	7	7.170375	1.024339	20.820
Error	8	0.393600	0.049200	

Then the F value is highly significant and we conclude that treatment means are different.

To take advantage of the factorial structure, we partition the treatment sum of squares into eight components, each with 1 degree of freedom. The main effect and interaction sums of squares with F values are given below:

Source*	df	SS	F
A	1	0.319225	6.488
B	1	1.288225	26.183
C	1	4.862025	98.822
AB	1	0.198025	4.025
AC	1	0.354025	7.196
BC	1	0.015625	0.318
ABC	1	0.133225	2.708

The tabulated F value at the 0.0005 level with 7 and 8 degrees of freedom is 15.1, so we conclude that treatments are different. The main effect and interaction sums of squares help us determine where the differences lie. Of course the factor C, exhaust level, is the greatest source of variation. At the 5% level, all three main effects and the AC interaction are significant. To understand the AC interaction, let us look at the following two-way table of means.

		C	
		13	35
A	85 min	1.7775	3.1775
	120 min	1.7925	2.5975

---

* A = Average irradiation time

B = Fuel composition

C = Exhaust level

At the lower exhaust level, the effect of increasing irradiation time is to increase the NO_2 formation rate. However at the higher exhaust level, the effect of increasing irradiation time is to decrease the NO_2 formation rate.

2. Completely Randomized Design with Eight Experimental Units

In physical experiments the type of duplication often used has been described as "bang-bang" duplication. That is a treatment which is run and then run again before going to another treatment. With this sort of approach, the randomization might have been as follows:

Run	Exhaust Level ppm carbon	Fuel Composition % elegins	Average Irradiation Time min
1	35	23	120
2	35	23	120
3	13	23	120
4	13	23	120
5	35	23	85
6	35	23	85
7	13	14	120
8	13	14	120
9	35	14	85
10	35	14	85
11	13	23	85
12	13	23	85
13	35	14	120
14	35	14	120
15	13	14	85
16	13	14	85

With the experiment run in this fashion, it appears to us that the experimental unit consists of the two successive time periods for which a treatment is used. Then we have only eight experimental units and zero degrees of freedom for experimental error. If we use the residual mean square in the analysis of variance for experimental error, we will declare more results significant than are justified by the data.

3. Randomized Block Design with 16 Experimental Units

Suppose that we run all eight treatments and then repeat the experiment. Then we have two blocks, consisting of eight runs each. Suppose the data were as follows:

Exhaust Level ppm carbon	Fuel Composition % elegins	Average Irrad. Time min.	NO_2, ppm/min Block I	NO_2, ppm/min Block II
13	14	85	1.25	1.76
13	14	120	1.45	1.67
13	23	85	1.88	2.22
13	23	120	1.95	2.10
35	14	85	2.42	2.90
35	14	120	2.36	2.61
35	23	85	3.59	3.80
35	23	120	2.68	2.74

Then the analysis of variance would be as follows:

AOV(3)

Source	df	SS	MS	F
Total	15	7.563975		
Blocks	1	0.308025		
Treatments	7	7.170375	1.024339	83.7905
Error	7	0.085575	0.012225	

With this analysis, all of the factorial effects and interactions, except the BC interaction, would be significant at the 5% level.

6.12 Mathematical Models

For planned experimentation, a general description of most situations is that one has a set of conditions, which can be more or less controlled, and a response measurable on a numeric scale, which can be made to vary by manipulation of the conditions. The objective, of course, is to be able to relate the response to the settings of the conditions in such a way as to permit predictions to be made about the response obtainable from a given combination of conditions. The most common approach is to use a mathematical model to approximate the relationship. We have already discussed the fact that exact reproducibility is never attained and have discussed how the use of probability theory can assist in characterizing non-reproducibility or random variation, as we have called it.

It is evident that the large selection of available mathematical functions and the great variety of probability distributions allow almost unlimited latitude in representing phenomena of nature with bi-component models. It is also evident that the more familiar and knowledgable about a particular phenomenon one becomes, the more representative and sophisticated he can make his model. One occasionally hears the comment that statistical methods should be avoided because they encourage empiricism. This criticism is not wholly without foundation, and it is often the result of the observation that a relatively crude mathematical model coupled with a probability component can sometimes do an amazingly good job of representing nature. This, of course, attests to the power of the tool of probability. However, the criticism is also applied to cases where a crude model has been used when a better one exists and is readily available. This is related to the matter of discriminating between systematic variation and random variation or, as we have previously called it, explainable variation versus unexplainable variation. We must agree that the fact that statistics provides a means of dealing with random or unexplained variation in no way excuses sloppiness or laziness in finding the best model that resources will allow for describing systematic variation. However, there are a couple of additional points one should consider before he attaches the label "empiricist " to a colleague. First, one must always ask the questions, "What is the purpose of the experiment for which the predictive model is sought; how exact a model is needed in relation to that purpose; what is the cost of developing a more refined model; and is such a refinement justifiable?" Secondly, empiricality is largely a matter of degree for, generally, no matter how "theoretical" or "fundamental" a model one derives for a given situation, there remain unknown constants or parameters to be determined by experimental observation. In that sense, all models are empirical.

From a statistical standpoint, models may be classified in several different ways. We may speak of purely stochastic models versus those which are not purely stochastic. A purely stochastic model is one which is based wholly upon one or more probability distributions; in other words, one which is made up of only the probability component. The most common type of model, however, is one which is composed of some closed function of the experimental conditions, with a random error added on ; i.e., we say

$$\text{Response} = f(\text{Experimental Conditions}) + (\text{Random Error}).$$

Of the models in this class, two subclasses are differentiated: *linear models* and *non-linear models*. To simplify the discussion, let us assume that the "experimental conditions" can be translated into a set of numerically measurable quantitative factors. This, of course, is the common situation in the physical sciences. More specifically, let us suppose that the set of variables $(x_1, x_2, \ldots, x_k)$ has been identified as one which should account for "most of" the variation in some response variable, say, y. Then, linearity refers to the form of the function f in the model

$$y = f(x_1, x_2, \ldots, x_k) + \text{error}.$$

A *linear model* is one for which the function f can be expressed as a linear combination of functions

$$f(x_1, x_2, \ldots, x_k) = b_1 f_1(x_1, x_2, \ldots, x_k) + b_2 f_2(x_1, x_2, \ldots, x_k) + \\ + b_m f_m(x_1, x_2, \ldots, x_k),$$

where $b_1, b_2, \ldots, b_n$ are constants to be determined by experimentation. Any model which cannot be put into this form is said to be *non-linear*. Let us note that a non-linear model may sometimes be linearized by transformation. For example, a model such as

$$y = b_0 x_1^{b_1} x_2^{b_2} [(x_3+x_4)/x_5]^{b_3}$$

may also be expressed in log units as

$$\log y = \log b_0 + b_1 \log x_1 + b_2 \log x_2 + b_3 \log[(x_3+x_4)/x_5]$$

or

$$y' = a_0 + a_1 z_1 + a_2 z_2 + b_3 z_3 ,$$

which is a linear model. For this reason, the term *intrinsically non-linear* is sometimes used to describe non-linear models which are not linearizable.

One may wonder why such distinctions as linear, linearizable, or intrinsically non-linear are made, since the general formulation is the same in any case. Let us note, however, that up to now in the discussion of models, there has been nothing covered which is uniquely associated with, or is a unique product of, statistical theory, other than the idea that one may include, as an integral part of any model, a probability component. From a statistical standpoint, it is possible as well as advantageous to view the models we have been discussing entirely as probability models; that is, to consider the response y as a random variable with its distribution determined by both the function f and the random error component. Looking at models in this way permits us to see the essentials involved, which are:

1. The determination (estimation) of unknown constants (parameters).
2. The determination of the behavior of the random component.
3. The determination of the behavior of the parameter estimates obtained in 1, based upon the information from 2.

While many scientists and engineers recognize the importance of 1 and 2, unfortunately a great many do not know that a suitably designed experiment can yield information for all three activities, rather than for 1 only. In order to keep one's perspective, it is helpful to be reminded that the activities we are discussing here are based upon the same basic principles covered in Chapter 5. That is, regardless of how complex the structure of the population (as described by some model) may be, our efforts, though they may be complex experiments, can be regarded as taking samples from that population. In Chapter 5, we saw that the inferences we draw to the population sampled depend upon what we assume about the structure of the population and how the population is sampled. It is no different in the present case; the interpretation of the results of an experiment is dependent upon the model chosen and the type of experiment conducted. Unfortunately, at this point, the distinction linear or non-linear becomes important for, given that the model is linear, it is a relatively straightforward matter to determine how an experiment should be conducted and analyzed so that objectives 1, 2, and 3 above may be met in an optimal manner. For this reason, statistical theory of model fitting has, until recently, concentrated upon the linear case, with the result that the non-linear case has lagged behind in the development of a unified statistical approach. The situation is being remedied, however, to the point that nowadays one need not compromise the usefulness of his research by settling for a linear model to describe a population structure when a non-linear representation is definitely required.

Since linear models have occupied a central role in the development of statistical approaches to the gathering and analysis of data, let us briefly discuss the essential ideas involved. Recall that a linear model is one of the form

$$y = b_1 f_1(x_1, x_2, \ldots, x_k) + b_2 f_2(x_1, x_2, \ldots, x_k) + \ldots + b_m f_m(x_1, x_2, \ldots, x_k)$$
$$+ \text{ error}.$$

Let $z_i = f_i(x_1, x_2, \ldots, x_k)$ and denote the error term by the letter e; then the model is

$$y = b_1 z_1 + b_2 z_2 + \ldots + b_m z_m + e .$$

Suppose that an experiment is conducted for which n observations $(y_1, y_2, \ldots, y_n)$ of the response are taken. Denoting the values of the conditions associated with y_i by $(z_{i1}, z_{i2}, \ldots, z_{im})$, we may express the n observations as

$$\begin{aligned} y_1 &= b_1z_{11} + b_2z_{12} + \cdots + b_mz_{1m} + e_1 \\ y_2 &= b_1z_{21} + b_2z_{22} + \cdots + b_mz_{2m} + e_2 \\ &\;\vdots \\ y_n &= b_1z_{n1} + b_2z_{n2} + \cdots + b_mz_{nm} + e_n \end{aligned} ,$$

where the b_i's and e_i's are unknown. Matrix notation may be used to express the situation more concisely by

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \cdots z_{1m} \\ z_{21} & z_{22} \cdots z_{2m} \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ z_{n1} & z_{n2} \cdots z_{nm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

or, simply

$$Y = X\beta + e .$$

Now, since the vector Y and the matrix X are determined respectively by the experimental observations and conditions, the vectors β and e are the only undetermined quantities in the expression. Also, the vector e we are regarding as the errors between the observed responses Y and those responses computed from the model. It is clear, of course, that the $m + n$ b_i's and e_i's cannot all be solved for uniquely with only n equations. Imagine for the moment that the b_i's are known and the e_i's are unknown, but that the e_i's are random fluctuations about zero. In such a case, we would say that the responses y_i, for the associated conditions $z_{i1}, z_{i2}, \ldots, z_{im}$, predicted by the model are

$$\hat{y}_i = b_1 z_{i1} + b_2 z_{i2} + \ldots + b_m z_{im}, \qquad i = 1, 2, \ldots, n \ ,$$

and we would call the respective e_i's the errors of prediction. The smaller the e_i's are as a group, the better, we say, that the model and data agree.

This, then, suggests a method of choosing the b_i's; that is, to find the set of b_i's that makes the discrepancies between y_i and $\hat{y}_i$ as small as possible. As a first cut, one might consider minimizing the sum of the unsigned differences $\sum_{i=1}^{n} |y_i - \hat{y}_i|$. A better way, from some points of view, is to minimize the sum of squared differences $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. This is, of course, the familiar method of least squares. The details of the method may be found in many statistical references, so let us simply summarize what is done. With the matrix notation formulation of the model given earlier, it can be shown that $\hat{\beta}$ is such that

$$X'X\hat{\beta} = X'Y$$

is satisfied. For those unfamiliar with matrix notation, the above is simply a shorthand way of saying that the b_i's must satisfy a system of simultaneous linear equations. This particular set of equations is called the normal equations. Thus, what it boils down to is that to utilize a linear model to describe the results of an experiment, the parameters to be determined by the method of least squares are found by solving the normal equations. In this day of readily accessible high speed computers, one usually lets a computer perform the laborious computations involved.

Let us now return to the idea that the model is determined for the purpose of predicting the responses one might expect to get from experimental conditions "similar" to those used to determine the model. Once again, one is confronted with the fact that, in all probability, even though identical conditions are ostensibly met at some future time, the responses observed will not agree with those predicted by any model and, in fact, will not agree with the responses obtained in the experiment being analyzed. The disagreement between the responses predicted from the chosen model and the observed responses is often used to measure the variability to be expected if the experiment were repeated indefinitely. One can see here a fundamental and nontrivial difficulty; namely, that the measure of variability depends upon the model used. In other words, the discrepancy between predicted and observed might be largely a result of the model being incorrect, sometimes referred to as lack of fit of the model. For that reason, many researchers, whenever possible, make it a standard practice to replicate experimental conditions. In the context of models, we may think of replication as obtaining several independent responses at each experimental condition. Then, as long as the factors included in the predictive model are constant, the several independent responses for the same "setting" would be predicted to be constant. Hence, the failure of the several observed responses to all be equal to the predicted response may be broken down into two components: (1) the failure of the observed responses to all be equal to their mean value, and (2) the failure of the mean value of the observed responses to be the same as the response predicted by the model. In this way, one can get respective measures of both the random variation and the lack of fit of the model. The result of not replicating is to force one to assume that lack of fit of the model is merely random error in order to get a measure of variability, an assumption which may or may not be justified. This is why replication is recommended whenever possible.

7. SUMMARY

A large portion of scientific activity can be loosely characterized as "attempting to discern the general pattern" or "trying to predict the behavior of some response under a given set of conditions." These objectives would be fairly easy to attain if it were not for the fact that variability seems to be one property that all physical material possesses. In recent years, there has been a growing tendency to formally recognize variability by incorporating a provision for it into the models used to characterize physical happenings. This, in turn, has led to the practice of visualizing numerical observations upon a phenomenon as having arisen from some conceptual population of "similar" observations that would occur under "the same conditions." The principles of Statistical Inference have shown, in addition, that the idea of a population can be extended even further by thinking in terms of populations of samples which could be generated by repeated sampling from the original population. This way of looking at observational data has been an extremely important development in data analysis, because it has helped scientists to realize that what one may infer from a set of data (sample) depends very much upon how it was obtained; that is, the procedure by which "similar" samples would be obtained. Almost everyone is familiar with the type of calculations which can be performed on a set of numbers, giving rise to things such as mean, standard deviation, median, and percentiles. Whenever interest lies only in the set of numbers itself, such quantities are helpful in providing summary descriptions. However, when a set of numbers is considered to be a sample from a large finite population or from an infinite population, sample quantities such as mean and standard deviation are of limited value in terms of the larger population unless it is possible to ascertain something about their behavior over repeated sampling. There are two factors which determine this behavior: (1) the sampling procedure, and (2) the structure of the sampled population. We have seen that when the selection of samples is allowed to be governed by the laws of chance, evaluation of the performance of estimates and formulation of tests and confidence statements are possible to a certain degree, even when the sampled population is not given a structure. In a great many cases, however, it is advantageous to assume that the sampled population has an underlying probability structure of distribution. Of all the distributions available, the most commonly used for this purpose is the normal distribution. Models for population structures can be quite varied. The simplest type is one of the form $Y = \mu + e$, where μ is a fixed constant and e is a random variable. Such a model is frequently used to represent repeated measurements, where μ is the true value and e is a random measurement. Note that if we give e a probability distribution with mean zero and variance σ^2, then Y is a random variable also, having mean μ and variance σ^2. A simple model of this type with $e \sim N(0, \sigma^2)$ would describe the structure of most of the continuous populations discussed in the context of statistical inference to this point. Populations which are sampled by conducting experiments are generally considered to have structures described by more complex models such as $Y = f(\text{Set of Conditions}, e)$ where f is some function and e is some random error.

8. REFERENCES

1. Anscombe, F.J. The Validity of Comparative Experiments. Jour. Roy. Stat. Soc. A, 61:181-211, 1948.

2. Bennett, C.A. and N.L. Franklin. Statistical Analysis in Chemistry and in the Chemical Industry, New York, Wiley, 1954.

3. Bowker, A.H. and G.J. Lieberman, Engineering Statistics. Englewood Cliffs, N.J., Prentice-Hall, 1959.

4. Brownlee, K.A. Industrial Experimentation. New York, Chemical Pub. Co. Inc., 1947.

5. Brownlee, K.A. Statistical Theory and Methodology in Science and Engineering. New York, Wiley, 1965.

6. Chemical Rubber Company. Handbook of Tables for Probability and Statistics. Cleveland, Chemical Rubber Co., 1968.

7. Chew, V. Experimental Designs in Industry. New York, Wiley, 1958.

8. Cochran, W.G. and M. Cox. Experimental Designs, Second Edition, New York, John Wiley, 1957.

9. Conover, W.J. Practical Nonparametric Statistics. New York, John Wiley, 1971.

10. Cox, D.R. Planning of Experiments. John Wiley, New York, 1958.

11. Davies, O.L., ed. The Design and Analysis of Industrial Experiments. New York, Hafner, 1954.

12. Davies, O.L., ed. Statistical Methods in Research and Production. New York, Hafner, 1957.

13. Draper, N.R. and H. Smith. Applied Regression Analysis. New York, Wiley, 1966.

14. Fisher, R.A. The Arrangement of Field Experiments. Journal of Ministry of Agric. 33:503-513, 1926

15. Fisher, R.A. The Design of Experiments. Edinburgh, Oliver, and Boyd, 1947.

16. Fisher, R.A. and W.A. MacKenzie. Studies in Crop Variation, II. The Manurial Response of Different Potato Varieties. Journal of Agricultural Science, 13:311, 1923.

17. Fisher, R.A. and F. Yates. Statistical Tables for Biological, Agricultural, and Medical Research, Sixth Edition. Edinburgh, Oliver, and Boyd, 1963.

18. Hald, A. Statistical Tables and Formulas. New York, John Wiley, 1952.

19. Hald, A. Statistical Theory with Engineering Applications. New York, Wiley, 1965.

20. Hogg, R.V. and A.T. Craig. Introduction to Mathematical Statistics, 2nd ed. New York, Macmillan, 1965.

21. Hooke. Introduction to Scientific Inference. San Francisco, Holden-Day, 1963.

22. Kempthorne, O. The Design and Analysis of Experiments. New York, John Wiley, 1952.

23. Kempthorne, O. and J.L. Folks. Probability, Statistics, and Data Analysis. Ames, Iowa, Iowa State Press, 1971.

24. Korth, W. Dynamic Irradiation Chamber Tests of Automotive Exhaust, Public Health Service Publication No. 999-AP-5, 1963.

25. Larsen, H.J. Introduction to Probability Theory and Statistical Inference. New York, Wiley, 1969.

26. Miller, I. and J.E. Freund. Probability and Statistics for Engineers. Englewood Cliffs, N.J., Prentice-Hall, 1965.

27. Mood, A.M. and F.A. Graybill. Introduction to the Theory of Statistics, 2nd ed. New York, McGraw-Hill, 1963.

28. Murphy, J.R. Procedures for Grouping a Set of Observed Means. Unpublished Dissertation. Oklahoma State University, 1973.

29. Ostle, B. Statistics in Research, 2nd ed. Ames, Iowa, Iowa State University Press, 1963.

30. Rand Corporation. A Million Random Digits with 100,000 Normal Deviates. Glencoe, Illinois, Free Press, 1955.

31. Siegel, Nonparametric Statistics for the Behavioral Sciences. New York, McGraw-Hill, 1956.

32. Snedecor. Statistical Methods, First Edition. Ames, Iowa, Iowa State Press, 1937.

33. Walpole, R.E. and R.H. Meyers. Probability and Statistics for Engineers and Scientists. New York, Macmillan, 1972.

34. Winer, B.J. Statistical Principles in Experimental Design, 2nd ed. New York, McGraw-Hill, 1971.

35. Youden, W.J. Statistical Methods for Chemists. New York, Wiley, 1951.